The American Economic Review

ARTICLES

JUNE 1987

THE AMERICAN ECONOMIC ASSOCIATION

•Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

•*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Survey*, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Regular member dues (nonrefundable) for 1987, which include a subscription to the *American Economic Review*, *Journal of Economic Literature*, and *Journal of Economic Perspectives*, are:

$38.50 if annual income is $30,000 or less;
$46.20 if annual income is above $30,000, but no more than $40,000;
$53.90 if annual income is above $40,000.

Nonmember subscriptions will be accepted only for all journals: Institutions (libraries, firms, etc.), $110 a year; individuals, $72.00. Single copies of journals may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $12.00 to the annual rates above to cover extra postage.

•Correspondence relating to the *Survey*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. **Change of address must reach the Secretary at least six (6) weeks prior to the month of publication.** The Association's publications are mailed second class. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

•Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

•Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

THE AMERICAN ECONOMIC REVIEW

•Submit manuscripts (4 copies), 50 pages maximum, double spaced, to:
Orley Ashenfelter, Managing Editor, *AER*; 209 Nassau Street, Princeton, NJ 08542-4607.

•Submission fee: $50 for members; $100 for nonmembers. Style guides will be provided upon request.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

June 1987

3̶91
, 001

VOLUME 77, NUMBER 3

Articles

Shorter Papers

THE AMERICAN ECONOMIC REVIEW

• Submit manuscripts (4 copies), 50 pages maximum, double spaced, to:
Orley Ashenfelter, Managing Editor, *AER*; 209 Nassau Street, Princeton, NJ 08542-4607.

• Submission fee: $50 for members; $100 for nonmembers. Style guides will be provided upon request.

June 1987

VOLUME 77, NUMBER 3

Articles

Shorter Papers

The Constitution of Economic Policy†

By JAMES M. BUCHANAN*

I. Introduction

> The science of public finance should always keep...political conditions clearly in mind. Instead of expecting guidance from a doctrine of taxation that is based on the political philosophy of by-gone ages, it should instead endeavor to unlock the mysteries of the spirit of progress and development.
> [Wicksell, p. 87][1]

On this of all occasions I should be remiss if I failed to acknowledge the influence of that great Swede, Knut Wicksell, on my own work, an influence without which I should not be making this presentation. Many of my contributions, and especially those in political economy and fiscal theory, might be described as varied reiterations, elaborations, and extensions of Wicksellian themes; this paper is no exception.

One of the most exciting intellectual moments of my career was my 1948 discovery of Wicksell's unknown and untranslated dissertation, *Finanztheoretische Untersuchungen* (1896), buried in the dusty stacks of Chicago's old Harper Library. Only the immediate postdissertation leisure of an academic novice allowed for the browsing that produced my own dramatic example of learning by serendipity. Wicksell's new principle of justice in taxation gave me a tremendous surge of self-confidence. Wicksell, who was an established figure in the history of economic ideas, challenged the orthodoxy of public finance theory along lines that were congenial with my own developing stream of critical consciousness. From that moment in Chicago, I took on the determination to make Wicksell's contribution known to a wider audience, and I commenced immediately a translation effort that took some time, and considerable help from Elizabeth Henderson, before final publication.

Stripped to its essentials, Wicksell's message was clear, elementary, and self-evident. Economists should cease proffering policy advice as if they were employed by a benevolent despot, and they should look to the structure within which political decisions are made. Armed with Wicksell, I, too, could dare to challenge the still-dominant orthodoxy in public finance and welfare economics. In a preliminary paper (1949), I called upon my fellow economists to postulate some model of the state, of politics, before proceeding to analyze the effects of alternative policy measures. I urged economists to look at the "constitution of economic policy," to examine the rules, the constraints within which political agents act. Like Wicksell, my purpose was ultimately normative rather than antiseptically scientific. I sought to make economic sense out of the relationship between the individual and the state before proceeding to advance policy nostrums.

Wicksell deserves the designation as the most important precursor of modern public choice theory because we find, in his 1896 dissertation, all three of the constitutive elements that provide the foundations of this theory: methodological individualism, *homo economicus*, and politics-as-exchange. I shall discuss these elements of analytical structure in the sections that follow. In Section V, I integrate these elements in a theory of eco-

†This is the lecture James Buchanan delivered in Stockholm, Sweden, December 8, 1986, when he received the Nobel Prize in Economic Science.

*Center for Public Choice, George Mason University, Fairfax, VA 22030. I am indebted to Robert Tollison, Viktor Vanberg, and Richard Wagner for helpful comments.

[1]This and subsequent citations are from Knut Wicksell, "A New Principle of Just Taxation," included in R. A. Musgrave and A. T. Peacock (1958, pp. 72–118). The more inclusive work from which this translated essay is taken is Wicksell, *Finanztheoretische Untersuchungen* (1896).

nomic policy. This theory is consistent with, builds upon, and systematically extends the traditionally accepted principles of Western liberal societies. The implied approach to institutional-constitutional reform continues, however, to be stubbornly resisted almost a century after Wicksell's seminal efforts. The individual's relation to the state, is, of course, the central subject matter of political philosophy. Any effort by economists to shed light on this relationship must be placed within this more comprehensive realm of discourse; a summary effort is contained in Section VI.

II. Methodological Individualism

> If utility is zero for each individual member of the community, the total utility for the community cannot be other than zero. [Wicksell, p. 77]

The economist rarely examines the presuppositions of the models with which he works. The economist simply commences with individuals as evaluating, choosing, and acting units. Regardless of the possible complexity of the processes or institutional structures from which outcomes emerge, the economist focuses on individual choices. In application to market or private-sector interactions, this procedure is seldom challenged. Individuals, as buyers and sellers of ordinary (legally tradable) goods and services are presumed able to choose in accordance with their own preferences, whatever these may be, and the economist does not feel himself obliged to inquire deeply into the content of these preferences (the arguments in individuals' utility functions). Individuals themselves are the sources of evaluation, and the economist's task is to offer an explanation-understanding of the process through which these unexamined preferences are ultimately translated into a complex outcome pattern.

The eighteenth-century discovery that, in an institutional framework that facilitates voluntary exchanges among individuals, this process generates results that might be evaluated positively, produced "economics," as an independent academic discipline or science. The relationship between the positively valued results of market processes and the institutional characteristics of these processes themselves emerged as a source of ambiguity when "the market" came to be interpreted functionally, as if something called "the economy" existed for the purpose of value maximization. Efficiency in the allocation of resources came to be defined independently of the processes through which individual choices are exercised.

Given this subtle shift toward a teleological interpretation of the economic process, it is not surprising that politics, or governmental process, was similarly interpreted. Furthermore, a teleological interpretation of politics had been, for centuries, the dominating thrust of political theory and political philosophy. The interpretations of "the economy" and "the polity" seemed, therefore, to be mutually compatible in the absence of inquiry into the fundamental difference in the point of evaluation. There was a failure to recognize that individuals who choose and act in the market generate outcomes that, under the specified constraints, can be judged to be value maximizing for participating individuals, *without* the necessity of introducing an external evaluative criterion. The nature of the process itself insures that individual values are maximized. This "value-maximization" perspective cannot be extended from the market to politics since the latter does not directly embody the incentive compatible structure of the former. There is no political counterpart to Adam Smith's invisible hand. It is not, therefore, surprising that the attempt by Wicksell and other continental European scholars to extend economic theory to the operation of the public sector remained undeveloped for so many years.

An economic theory that remains essentially individualistic need not have become trapped in such a methodological straight jacket. If the maximization exercise is restricted to explanation-understanding of the individual who makes choices, and without extension to the economy as an aggregation, there is no difficulty at all in analyzing individual choice behavior under differing institutional settings, and in predicting how these varying settings will influence the out-

comes of the interaction processes. The individual who chooses between apples and oranges remains the same person who chooses between the levers marked "Candidate A" and "Candidate B" in the polling booth. Clearly, the differing institutional structures may, themselves, affect choice behavior. Much of modern public choice theory explains these relationships. But my point here is the more basic one to the effect that the choice behavior of the individual is equally subject to the application of analysis in all choice environments. Comparative analysis should allow for predictions of possible differences in the characteristics of the results that emerge from market and political structures of interaction. These predictions, as well as the analysis from which they are generated, are totally devoid of normative content.

III. *Homo Economicus*

> ...[N]either the executive nor the legislative body, and even less the deciding majority in the latter, are in reality...what the ruling theory tells us they should be. They are not pure organs of the community with no thought other than to promote the common weal.
>
> ...[M]embers of the representative body are, in the overwhelming majority of cases, precisely as interested in the general welfare as are their constituents, neither more nor less.
>
> [Wicksell, pp. 86, 87]

This analysis can yield a limited set of potentially falsifiable hypotheses without prior specification of the arguments in individual utility functions. If, however, predictions are sought concerning the effects of shifts in constraints on choice behavior, some identification and signing of these arguments must be made. With this step, more extensive falsifiable propositions may be advanced. For example, if both apples and oranges are positively valued "goods," then, if the price of apples falls relative to that of oranges, more apples will be purchased relative to oranges; if income is a positively valued "good," and, then, if the marginal rate of tax on income source A increases relative to that on income source B, more effort at earning income will be shifted to source B; if charitable giving is a positively valued "good," then, if charitable gifts are made tax deductible, more giving will be predicted to occur; if pecuniary rents are positively valued, then, if a political agent's discretionary power to distribute rents increases, individuals hoping to secure these rents will invest more resources in attempts to influence the agent's decisions. Note that the identification and signing of the arguments in the utility functions takes us a considerable way toward operationalization without prior specification of the relative weights of the separate arguments. There is no need to assign net wealth or net income a dominating motivational influence on behavior in order to produce a fully operational economic theory of choice behavior, in market or political interaction.

In any extension of the model of individual rational behavior to politics, this difference between the identification and signing of arguments on the one hand and the weighting of these arguments on the other deserves further attention. Many critics of the "economic theory of politics" base their criticisms on the presumption that such theory necessarily embodies the hypothesis of net wealth maximization, an hypothesis that they observe to be falsified in many situations. Overly zealous users of this theory may have sometimes offered grounds for such misinterpretation on the part of critics. The minimal critical assumption for the explanatory power of the economic theory of politics is only that identifiable economic self-interest (for example, net wealth, income, social position) is a positively valued "good" to the individual who chooses. This assumption does not place economic interest in a dominating position and it surely does not imply imputing evil or malicious motives to political actors; in this respect the theory remains on all fours with the motivational structure of the standard economic theory of market behavior. The differences in the predicted results stemming from market and political interaction stem from differences in

the structures of these two institutional settings rather than from any switch in the motives of persons as they move between institutional roles.

IV. Politics as Exchange

> It would seem to be a blatant injustice if someone should be forced to contribute toward the costs of some activity which does not further his interests or may even be diametrically opposed to them. [Wicksell, p. 89]

Individuals choose, and as they do so, identifiable economic interest is one of the "goods" that they value positively, whether behavior takes place in markets or in politics. But markets are institutions of *exchange*; persons enter markets to exchange one thing for another. They do not enter markets to further some supra-exchange or supra-individualistic result. Markets are not motivationally functional; there is no conscious sense on the part of individual choosers that some preferred aggregate outcome, some overall "allocation" or "distribution," will emerge from the process.

The extension of this exchange conceptualization to politics counters the classical prejudice that persons participate in politics through some common search for the good, the true, and the beautiful, with these ideals being defined independently of the values of the participants as these might or might not be expressed by behavior. Politics, in this vision of political philosophy, is instrumental to the furtherance of these larger goals.

Wicksell, who is followed in this respect by modern public choice theorists, would have none of this. The relevant difference between markets and politics does not lie in the kinds of values/interest that persons pursue, but in the conditions under which they pursue their various interests. Politics is a structure of complex exchange among individuals, a structure within which persons seek to secure collectively their own privately defined objectives that cannot be efficiently secured through simple market exchanges. In the absence of individual interest, there is no interest. In the market, individuals exchange apples for oranges; in politics, individuals exchange agreed-on shares in contributions toward the costs of that which is commonly desired, from the services of the local fire station to that of the judge.

This ultimately voluntary basis for political agreement also counters the emphasis on politics as power that characterizes much modern analysis. The observed presence of coercive elements in the activity of the state seems difficult to reconcile with the model of voluntary exchange among individuals. We may, however, ask: Coercion to what purpose? Why must individuals subject themselves to the coercion inherent in collective action? The answer is evident. Individuals acquiesce in the coercion of the state, of politics, only if the ultimate constitutional "exchange" furthers their interests. Without some model of exchange, no coercion of the individual by the state is consistent with the individualistic value norm upon which a liberal social order is grounded.

V. The Constitution of Economic Policy

> ...[W]hether the benefits of the proposed activity to the individual citizens would be greater than its cost to them, no one can judge this better than the individuals themselves. [Wicksell, p. 79]

The exchange conceptualization of politics is important in the derivation of a normative theory of economic policy. Improvement in the workings of politics is measured in terms of the satisfaction of that which is desired by individuals, whatever this may be, rather than in terms of moving closer to some externally defined, supra-individualistic ideal. That which is desired by individuals may, of course, be common for many persons, and, indeed, the difference between market exchange and political exchange lies in the sharing of objectives in the latter. The idealized agreement on the objectives of politics does not, however, allow for any supersession of individual evaluation. Agreement itself emerges, again conceptually, from the revealed choice behavior of individuals. Commonly shared agreement must be carefully distinguished from any externally de-

fined definition or description of that "good" upon which persons "should agree."

The restrictive implications for a normative theory of economic policy are severe. There is no criterion through which policy may be directly evaluated. An indirect evaluation may be based on some measure of the degree to which the political process facilitates the translation of expressed individual preferences into observed political outcomes. The focus of evaluative attention becomes the process itself, as contrasted with end-state or outcome patterns. "Improvement" must, therefore, be sought in reforms in process, in institutional change that will allow the operation of politics to mirror more accurately that set of results that are preferred by those who participate. One way of stating the difference between the Wicksellian approach and that which is still orthodoxy in normative economics is to say that the *constitution* of policy rather than policy itself becomes the relevant object for reform. A simple game analogy illustrates the difference here. The Wicksellian approach concentrates on reform in the rules, which may be in the potential interest of *all* players, as opposed to improvement in strategies of play for particular players within defined or existing rules.

In the standard theory of choice in markets, there is little or no concern with the constitution of the choice environment. We simply presume that the individual is able to implement his preferences; if he wants to purchase an orange, we presume that he can do so. There is no institutional barrier between the revealed expression of preference and direct satisfaction. Breakdown or failure in the market emerges, not in the translation of individual preferences into outcomes, but in the possible presentation of some choosers with alternatives that do not correspond to those faced by others in the exchange nexus. "Efficiency" in market interaction is insured if the participants are faced with the same choice options.

In political exchange, there is no decentralized process that allows "efficiency" to be evaluated deontologically, akin to the evaluation of a market. Individuals cannot, by the nature of the goods that are collectively "purchased" in politics, adjust their own behavior to common terms of trade. The political analogue to decentralized trading among individuals must be that feature common over all exchanges, which is *agreement* among the individuals who participate. The unanimity rule for collective choice is the political analogue to freedom of exchange of partitionable goods in markets.

It is possible, therefore, to evaluate politics independently of results only by ascertaining the degree of correspondence between the rules of reaching decisions and the unique rule that would guarantee "efficiency," that of unanimity or agreement among all participants. If, then, "efficiency" is acknowledged to be the desired criterion, again as interpreted here, normative improvement in process is measured by movement toward the unanimity requirement. It is perhaps useful to note, at this point, that Wicksell's own characterization of his proposals in terms of "justice" rather than "efficiency" suggests the precise correspondence of these two norms in the context of voluntary exchange.

Politics as observed remains, of course, far from the idealized collective-cooperative exchange that the unanimity rule would implement. The political equivalent to transactions cost makes the pursuit of idealized "efficiency" seem even more out of the bounds of reason than the analogous pursuit in markets. But barriers to realization of the ideal do not imply rejection of the benchmark definition of the ideal itself. Instead, such barriers are themselves incorporated into a generalized "calculus of consent."

Wicksell himself did not go beyond advocacy of reform in legislative decision structures. He proposed a required linking of spending and financing decisions, and he proposed that a quasi-unanimity rule be introduced for noncommitted outlays. Wicksell did not consciously extend his analysis to constitutional choice, to the choice of the rules within which ordinary politics is to be allowed to operate. His suggested reforms were, of course, constitutional, since they were aimed to improve the process of decision making. But his evaluative criterion was restricted to the matching of individual preferences with political outcomes in par-

ticularized decisions, rather than over any sequence.

It is perhaps worth noting that Wicksell himself did not look upon his suggested procedural reforms as restrictive. By introducing greater flexibility into the tax-share structure, Wicksell predicted the potential approval of spending programs that would continue to be rejected under rigid taxing arrangements. Critics have, however, interpreted the Wicksellian unanimity constraint to be restrictive, and especially as compared to the extended activity observed in ordinary politics. This restrictive interpretation was perhaps partially responsible for the continued failure of political economists to recognize his seminal extension of the efficiency norm to the political sector. Such restrictiveness is very substantially reduced, and, in the limit, may be altogether eliminated, when the unanimity criterion is shifted one stage upward, to the level of potential agreement on constitutional rules within which ordinary politics is to be allowed to operate. In this framework, an individual may rationally prefer a rule that will, on particular occasions, operate to produce results that are opposed to his own interests. The individual will do so if he predicts that, on balance over the whole sequence of "plays," his own interests will be more effectively served than by the more restrictive application of the Wicksellian requirement in-period. The in-period Wicksellian criterion remains valid as a measure of the particularized efficiency of the single decision examined. But the in-period violation of the criterion does not imply the inefficiency of the rule so long as the latter is itself selected by a constitutional rule of unanimity.[2]

As noted, the shift of the Wicksellian criterion to the constitutional stage of choice among rules also serves to facilitate agreement, and, in the limiting case, may remove altogether potential conflicts among separate individual and group interests. To the extent that the individual reckons that a constitutional rule will remain applicable over a long sequence of periods, with many in-period choices to be made, he is necessarily placed behind a partial "veil of uncertainty" concerning the effects of any rule on his own predicted interests. Choice among rules will, therefore, tend to be based on generalizable criteria of fairness, making agreement more likely to occur than when separable interests are more easily identifiable.

The political economist who operates from within the Wicksellian research program, as modified, and who seeks to offer normative advice must, of necessity, concentrate on the process or structure within which political decisions are observed to be made. Existing constitutions, or structures of rules, are the subject of critical scrutiny. The conjectural question becomes: Could these rules have emerged from agreement by participants in an authentic constitutional convention? Even here, the normative advice that is possible must be severely circumscribed. There is no external set of norms that provides a basis for criticism. But the potential economist may, cautiously, suggest changes in procedures, in rules, that may come to command general assent. Any suggested change must be offered only in the provisional sense, and, importantly, it must be accompanied by a responsible recognition of political reality. Those rules and rules changes worthy of consideration are those that are predicted to be workable within the politics inhabited by ordinary men and women, and not those that are appropriate only for idealized, omniscient, and benevolent beings. Policy options must remain within the realm of the feasible, and the interests of political agents must be recognized as constraints on the possible.

VI. Constitutionalism and Contractarianism

> The ultimate goal...is equality before the law, greatest possible liberty, and the economic well-being and peaceful cooperation of all people.
>
> [Wicksell, p. 88]

[2] In my own retrospective interpretation, the shift of the Wicksellian construction to the constitutional stage of choice was the most important contribution in *The Calculus of Consent* (1962), written jointly with Gordon Tullock.

As the basic Wicksellian construction is shifted to the choice among rules or constitutions and as a veil of uncertainty is utilized to facilitate the potential bridging of the difference between identifiable and general interest, the research program in political economy merges into that of contractarian political philosophy, both in its classical and modern variations. In particular, my own approach has affinities with the familiar construction of John Rawls (1971), who utilizes the veil of ignorance along with the fairness criterion to derive principles of justice that emerge from a conceptual agreement at a stage prior to the selection of a political constitution.

Because of his failure to shift his own analytical construction to the level of constitutional choice, Wicksell was confined to evaluation of the political process in generating current allocative decisions. He was unable, as he quite explicitly acknowledged, to evaluate political action involving either prior commitments of the state, for example, the financing of interest on public debt, or fiscally implemented transfers of incomes and wealth among persons and groups. Distributional questions remain outside the Wicksellian evaluative exercise, and because they do so, we locate another source of the long-continued and curious neglect of the fundamental analytical contribution. With the shift to the constitutional stage of politics, however, this constraint is at least partially removed. Behind a sufficiently thick veil of uncertainty and/or ignorance, contractual agreement on rules that allow for some in-period fiscal transfers seems clearly to be possible. The precise features of a constitutionally approved transfer structure cannot, of course, be derived independently because of the restriction of evaluative judgment to the process of constitutional agreement. In this respect, the application is fully analogous to Wicksell's unwillingness to lay down specific norms for tax sharing independently of the process of agreement. *Any* distribution of tax shares generating revenues sufficient to finance the relevant spending project passes Wicksell's test, provided only that it meets with general agreement. Analogously, *any* set of arrangements for implementing fiscal transfers, in-period, meets the constitutional stage Wicksellian test, provided only that it commands general agreement.

This basic indeterminacy is disturbing to political economists or philosophers who seek to be able to offer substantive advice, over and beyond the procedural limits suggested. The constructivist urge to assume a role as social engineer, to suggest policy reforms that "should" or "should not" be made, independently of any revelation of individuals' preferences through the political process, has simply proved too strong for many to resist. The scientific integrity dictated by consistent reliance on individualistic values has not been a mark of modern political economy.

The difficulty of maintaining such integrity is accentuated by the failure to distinguish explanatory and justificatory argument, a failure that has described the position of almost all critics of social contract theories of political order. We do not, of course, observe the process of reaching agreement on constitutional rules, and the origins of the rules that are in existence at any particular time and in any particular polity cannot satisfactorily be explained by the contractarian model. The purpose of the contractarian exercise is not explanatory in this sense. It is, by contrast, justificatory in that it offers a basis for normative evaluation. Could the observed rules that constrain the activity of ordinary politics have emerged from agreement in constitutional contract? To the extent that this question can be affirmatively answered, we have established a legitimating linkage between the individual and the state. To the extent that the question prompts a negative response, we have a basis for normative criticism of the existing order, and a criterion for advancing proposals for constitutional reform.[3]

It is at this point, and this point only, that the political economist who seeks to remain within the normative constraints imposed by

[3]A generalized argument for adopting the constitutionalist-contractarian perspective, in both positive and normative analysis, is developed in *The Reason of Rules* (1985), written jointly with Geoffrey Brennan.

the individualistic canon may enter the ongoing dialogue on constitutional policy. The deficit-financing regimes in modern Western democratic polities offer the most dramatic example. It is almost impossible to construct a contractual calculus in which representatives of separate generations would agree to allow majorities in a single generation to finance currently enjoyed public consumption through the issue of public debt that insures the imposition of utility losses on later generations of taxpayers. The same conclusion applies to the implicit debt obligations that are reflected in many of the intergenerational transfer programs characteristic of the modern welfare state.

The whole contractarian exercise remains empty if the critical dependence of politically generated results upon the rules that constrain political action is denied. If end states are invariant over shifts in constitutional structure, there is no role for constitutional political economy. On the other hand, if institutions do indeed matter, the role is well defined. Positively, this role involves analysis of the working properties of alternative sets of constraining rules. In a game-theoretic analogy, this analysis is the search for solutions of games, as the latter are defined by sets of rules. Normatively, the task for the constitutional political economist is to assist individuals, as citizens who ultimately control their own social order, in their continuing search for those rules of the political game that will best serve their purposes, whatever these might be.

In 1987, the United States celebrates the bicentennial anniversary of the constitutional convention that provided the basic rules for the American political order. This convention was one of the very few historical examples in which political rules were deliberately chosen. The vision of politics that informed the thinking of James Madison was not dissimilar, in its essentials, from that which informed Knut Wicksell's less comprehensive, but more focused, analysis of taxation and spending. Both rejected any organic conception of the state as superior in wisdom to the individuals who are its members. Both sought to bring all available scientific analysis to bear in helping to resolve the continuing question of social order: How can we live together in peace, prosperity, and harmony, while retaining our liberties as autonomous individuals who can, and must, create our own values?

REFERENCES

Brennan, Geoffrey and Buchanan, James, *The Reason of Rules*, Cambridge: Cambridge University Press, 1985.

Buchanan, James M., "The Pure Theory of Public Finance: A Suggested Approach," *Journal of Political Economy*, December 1949, *57*, 496–505.

______ **and Tullock, Gordon,** *The Calculus of Consent*, Ann Arbor: University of Michigan Press, 1962.

Musgrave, R. A. and Peacock, A. T., *Classics in the Theory of Public Finance*, London: Macmillan, 1958.

Rawls, John, *A Theory of Justice*, Cambridge: Harvard University Press, 1971.

Wicksell, Knut, *Finanztheoretische Untersuchungen*, Jena: Gustav Fisher, 1896.

Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment

By WILLARD G. MANNING, JOSEPH P. NEWHOUSE, NAIHUA DUAN, EMMETT B. KEELER, ARLEEN LEIBOWITZ, AND M. SUSAN MARQUIS*

We estimate how cost sharing, the portion of the bill the patient pays, affects the demand for medical services. The data come from a randomized experiment. A catastrophic insurance plan reduces expenditures 31 percent relative to zero out-of-pocket price. The price elasticity is approximately −0.2. We reject the hypothesis that less favorable coverage of outpatient services increases total expenditure (for example, by deterring preventive care or inducing hospitalization).

Over the past four decades medical care costs have grown about 4 percent per year in real terms, and the share of GNP devoted to medical care has increased from 4.4 percent in 1950 to 10.7 percent in 1985 (Daniel Waldo, Katherine Levit, and Helen Lazenby, 1986). A prominent explanation of this rapid increase has emphasized the spread of health insurance, which has generated demand for both a higher quality and an increased quantity of medical services (Martin Feldstein, 1971, 1977). In turn, the spread of health insurance has been linked to the exemption of employer-paid health insurance premiums from the individual income tax (Feldstein and Elizabeth Allison, 1974; Feldstein and Bernard Friedman, 1977; Mark Pauly, 1986). Thus, the increase in expenditure is often portrayed as a type of market failure induced by public policy, although such an argument is not universally accepted (Morris Barer, Robert Evans, and Gregory Stoddart, 1979; Robert Evans, 1984; John Goddeeris and Burton Weisbrod, 1985).

No one has shown, however, that the spread of health insurance can quantitatively account for most of the sustained rise in health expenditure (Pauly, 1986). If it cannot, the widespread presumption that distorted prices (because of insurance) are inducing excess resources in medical care is not necessarily correct. Central to appraising the quantitative role of insurance, of course, is the magnitude of the demand response to changes in insurance. The literature exhibits substantial disagreement, by a factor of 10 or more, about the price elasticity, or coinsurance elasticity, of demand (Richard Rosett and Lien Fu Huang, 1973; Karen Davis and Louise Russell, 1972; Charles Phelps and Newhouse, 1974; Fred Goldman and Michael Grossman, 1978; Ann Colle and Grossman, 1978; Newhouse and Phelps, 1974, 1976).[1]

Such disagreement is not surprising in light of the problems of using nonexperimental data to estimate elasticities (Newhouse,

*The Rand Corporation, 1700 Main Street, Santa Monica, CA 90406. This research was conducted under grant 016B80 from the Department of Health and Human Services. Neither the Department nor the Rand Corporation necessarily agrees with the conclusions presented here. We acknowledge the comments of three very detailed anonymous referee reports. We, of course, are responsible for any remaining errors. Bernadette Benjamin and Jack Zwanziger provided invaluable programming support and assistance throughout the analysis underlying this paper. Individuals too numerous to mention provided advice, assistance, and support throughout the project; however, two of our government project officers, Larry Orr and James Schuttinga, deserve special mention; so too do our former colleagues Charles Phelps and Carl Morris.

[1]The elasticity estimates at the mean vary from around −0.1 to −2.1.

Phelps, and Marquis, 1980). In cross-sectional data, insurance is endogenous; those who expect to demand more services have a clear incentive to obtain more complete insurance, either by selecting a more generous option at the place of employment, by working for an employer with a generous insurance plan, or by purchasing privately more generous coverage.

Ignoring this selection issue (i.e., treating insurance as exogenous) has generally produced results showing that demand for medical care responds to insurance-induced variation in price. Treating insurance as endogenous, however, has generally led to coefficients with confidence intervals that are insignificantly different from zero at conventional levels (Newhouse and Phelps, 1976).[2]

That upward bias may be present is suggested by results from several natural experiments that compared demands of the same individuals before and after their group insurance changed (Anne Scitovsky and Nelda Snyder, 1972; Scitovsky and Nelda McCall, 1977; Phelps and Newhouse, 1972; R. G. Beck, 1974). In these cases the change in insurance is presumptively exogenous, and the elasticity estimates cluster near the low end of those cited above. But natural experiments have no control group, so that any other factor that changed over time is perfectly confounded with the insurance change. Moreover, the samples available in such studies are not necessarily representative of the general population, and the changes in insurance that could be studied were limited to those that occurred in the natural experiment. Hence, these results too have been suspect.[3]

In light of the uncertainty about how demand responds to insurance-induced changes in price, and the importance for both public and private decisions of quantifying that response, the federal government initiated the Rand Health Insurance Experiment (HIE) in 1974, one aim of which was to narrow uncertainty about this issue (Newhouse, 1974). In this article we report the results of that experiment. Our findings have implications for the role of insurance in explaining the postwar increase in medical expenditure, as well as for the magnitude of the welfare loss from health insurance.

The HIE had several objectives other than improved estimates of how demand responds to insurance. Four such objectives merit mention here:

1) Many poor individuals are insured through public programs; whether the demand response differs for the poor is therefore an issue in decisions on the scope of these programs.

2) Insurance need not be uniform across various medical services. In fact, second-best pricing implies that coverage should be more generous for less price elastic (or less insurance elastic) services (Frank Ramsey, 1927; Richard Zeckhauser, 1970; William Baumol and David Bradford, 1970). We therefore wished to learn if insurance elasticities differed for various types of medical services. In particular, are demand elasticities greater for outpatient physician services, psychotherapy, and preventive services, which would accord with the observed lesser coverage of these services?[4]

3) The public financing of medical care has been justified by its status as a merit good (Richard Musgrave, 1959) and in particular the claim that the consumption of medical services leads to improved health, which can generate externalities (Cotton Lindsay, 1969; Anthony Culyer, 1971, 1976, 1978; Pauly, 1971; Evans, 1984). Thus, we

[2]Although many believe this failure to reject the null hypothesis when insurance is treated as endogenous occurs because the insurance variable is only weakly identified, the magnitude of any upward bias in elasticity estimates from treating insurance as exogenous remains unknown. Hausman (or Wu) type tests have not been used to test for endogeneity, but if they failed to reject the null hypothesis of exogeneity, it could be for lack of power because of a lack of a useful set of instruments.

[3]For reviews of the nonexperimental demand literature and a discussion of its methodological problems, see Newhouse (1978; 1981).

[4]Other explanations, not mutually exclusive, for the lower coverage of these services include greater loading charges and asymmetric information between insurer and insured.

sought to quantify how the change in the consumption of medical services at the margin might affect health. The answer to this question would inform the political debate about the benefits of public financing of medical care services for the indigent and would also inform the insurance decisions of private agents such as employers and unions.

4) For the past decade, public policy has promoted Health Maintenance Organizations (HMOs) on the groups that such organizations were more efficient in the delivery of services. Almost all evidence of lower cost, however, came from uncontrolled settings, leaving unresolved the question of whether selection of healthier members or more efficient treatment was responsible for lower costs in HMOs (Harold Luft, 1981). Also unresolved was the question of whether any true reduction in services at HMOs might adversely affect health status. Therefore, we sought to decompose the observed lower use of services at one HMO into the pure effect of the HMO, on the one hand, and treating a possibly less sickly group of enrollees on the other. Moreover, we sought to determine whether any reduced use of services affected health status and satisfaction.

This article considers the first two questions in some detail and summarizes the findings on the latter two.

I. Data and Sample

A. *The Design of the Rand Health Insurance Experiment*[5]

Between November 1974 and February 1977, the HIE enrolled families in six sites: Dayton, Ohio; Seattle, Washington; Fitchburg, Massachusetts; Franklin County, Massachusetts; Charleston, South Carolina; and Georgetown County, South Carolina.[6]

Families participating in the experiment were assigned to one of 14 different fee-for-service insurance plans or to a prepaid group practice; additionally, some members already enrolled in the prepaid group practice were enrolled as a separate group. The fee-for-service insurance plans, the main focus of this article, had different levels of cost sharing, which varied over two dimensions: the coinsurance rate (percentage paid out-of-pocket) and an upper limit on annual out-of-pocket expenses. The coinsurance rates were 0, 25, 50, or 95 percent. Each plan had an upper limit (the Maximum Dollar Expenditure or MDE) on annual out-of-pocket expenses of 5, 10, or 15 percent of family income, up to a maximum of $1,000.[7] Beyond the MDE, the insurance plan reimbursed all covered expenses in full.

Covered expenses included virtually all medical services.[8] One plan had different coinsurance rates for inpatient and ambulatory medical services (25 percent) than for dental and ambulatory mental health services (50 percent). And on one plan, the families faced a 95 percent coinsurance rate for outpatient services, subject to a $150 annual limit on out-of-pocket expenses per person ($450 per family). In this plan, all inpatient services were free; in effect, this plan had approximately an outpatient individual deductible.[9]

[5]Newhouse (1974) and Robert Brook et al. (1979), provide fuller descriptions of the design. Newhouse et al. (1979) discuss the measurement issues for the second generation of social experiments, to which the HIE belongs. John Ware et al. (1980) discuss many aspects of data collection and measurement for health status.

[6]The sites were selected to represent the four census regions; to represent the range of city sizes (a proxy for the complexity of the medical delivery system); to cover a range of waiting times to appointment and physician per capita ratios (to test for the sensitivity of demand elasticities to nonprice rationing); and to include both urban and rural sites in the North and the South.

[7]The maximum was $750 in some site-years for the 25 percent coinsurance plans. The $1000 was kept fixed in nominal dollars from 1974 to 1981. During this time the medical care component of the CPI rose by 96 percent.

[8]See Lorraine Clasquin (1973) for a discussion of the rationale for the HIE structure of benefits. Nonpreventive orthodontia and cosmetic surgery (related to preexisting conditions) were not covered. Also excluded were outpatient psychotherapy services in excess of 52 visits per year per person. In the case of each exclusion, it is questionable whether anything could have been learned about steady-state demand during the 3- to 5-year lifetime of the experiment.

[9]The coinsurance rate for the 95 percent and individual deductible plans was actually 100 percent in the first

Families were assigned to these insurance plans using the Finite Selection Model (Carl Morris, 1979). This model was used to achieve as much balance across plans as possible while retaining randomization; that is, it minimizes the correlation between the experimental treatments and health, demographic, and economic covariates.

To study methods effects, the HIE employed four randomized subexperiments (Newhouse et al., 1979). We describe two here. To test for transitory demand effects (Charles Metcalf, 1973; Kenneth Arrow, 1975), 70 percent of the households were enrolled for three years; the remainder for five years. Also, to ensure that no one was worse off financially from participating in the study, families were paid a lump sum payment.[10] To test for a possible stimulus to utilization, 40 percent of the families were given an unanticipated increase in their lump sum payment during the next to the last year of the study.

B. *The Sample*

The enrolled sample is for the most part a random sample of each site's nonaged population, but some groups were not eligible.[11] Table 1 gives the sample by plan and site; it excludes the 1,982 persons in the HMO experiment. Note that plans are not perfectly balanced by site; in particular, no one was enrolled in the 50 percent plan in Seattle, and about half of those in the 50 percent plan are in Dayton, whereas only 20 percent of all participants are in Dayton.[12]

1. *Dependent Variables*. In the interest of brevity, we focus primarily on the use of medical services other than outpatient psychotherapy and dental services.[13] We do, however, summarize results for dental services below.

year of Dayton, the first site. The rate was changed to 95 percent for all other site-years of the experiment, in order to increase the incentive to file claims, although there was no statistical evidence at that time of underfiling. Subsequent analysis has shown that the mean outpatient physician expenditure on the 95 percent coinsurance plans relative to the free-care plan is understated by about 5 to 10 percent because of a lower propensity to file claims (William Rogers and Newhouse, 1985).

[10]Because of size of the lump sum payment, there is a theoretical presumption of no bias from refusal or attrition. Although refusal and attrition occurred at higher rates on higher coinsurance plans, refusal and attrition appear to have been random within plan. More precisely, we detect no differences by plan at enrollment in pre-experimental use or health status, nor do we detect differences in the rate of spending between those who withdrew from the experiment and those who did not. More detailed data on issues of refusal and attrition can be found in Brook et al. (1983, 1984); Kevin O'Grady et al. (1985); Newhouse et al. (1987). The details of the lump sum payment rules can be found in Clasquin and Marie Brown (1977).

[11]The ineligible groups include: 1) those 62 years of age and older at the time of enrollment; 2) those with incomes in excess of $25,000 in 1973 dollars or $58,000 in 1984 dollars); this excluded 3 percent of the families contacted; 3) those eligible for the Medicare disability program, 4) those in jails or institutionalized for indefinite periods; 5) those in the military or their dependents; and 6) veterans with service-connected disabilities.

[12]About 3 percent of the actual participant-years are truncated because the participant withdrew partway through an accounting year. With the exception of deaths, we do not use such participants in the estimation sample because the 4-part model (see below) requires equal time periods for each observation. If a person is only observed for one quarter and the expenditure distribution is lognormal, the annual distribution is not simply the quarterly distribution scaled up by a factor of 4; i.e., the lognormal distribution does not convolute. The sample used in this analysis more specifically includes enrollees during each full year that they participated, and the last accounting year in the study for those who died. We excluded data on partial years of participation by newborns. (Their expenses in the hospital at the time of birth, however, are attributed to the mother.) We tested the legitimacy of excluding those with partial years by comparing expenditure rates of part-year persons, adjusted for time at risk, with what they would have spent if they behaved like full-year people. Specifically we regressed actual expenditure minus (time at risk times the 4-part model prediction) on plan dummy variables. We could not reject the null hypothesis of no difference by plan ($\chi^2(4) = 2.67$, $p > .50$). The estimated effect of including part-year participants is to negligibly increase the estimated response to plan.

[13]See Manning et al. (1984b, 1986b) and Kenneth Wells et al. (1982) for additional results on the use of mental health care, and Manning et al. (1985) for additional results on dental use. Mental health care use is on the order of 4 percent of the expenditures discussed here.

TABLE 1—NUMBER OF PERSONS AT ENROLLMENT AND NUMBER OF PERSON-YEARS IN ESTIMATION SAMPLE

	Site							
Plan	Dayton	Seattle	Fitch-burg	Frank-lin County	Charles-ton	George-town	Enroll-ment Total[a]	Esti-mation Sample Total[b]
Free	301	431	241	297	264	359	1893	6822
25 Percent[c]	260	253	125	152	146	201	1137	4065
50 Percent	191	0	56	58	26	52	383	1401
95 Percent	280	253	113	162	146	166	1120	3727
Individual Deductible	105	285	188	220	196	282	1276	4175
Total	1137	1222	723	889	778	1060	5809	20190

[a] Persons.
[b] Person-years.
[c] Includes those with 50 percent coinsurance for dental and mental health and 25 precent coinsurance for all other services.

2. *Independent Variables.* Although we present sample means by plan, we also present results controlling for site, health status, sociodemographic, and economic variables.

Insurance Plan Variables. Rather than impose a functional form, we have conservatively used dummy variables for insurance plans. We have grouped the insurance plans into five groups: 1) the free plan (no out-of-pocket cost to the family); 2) 25 percent coinsurance rate plans for medical services; 3) 50 percent coinsurance rate plans for medical services; 4) 95 percent coinsurance rate plans for medical services; and 5) the plan with a 95 percent coinsurance rate for outpatient services (subject to a limit of $150 per person or $450 per family per year) and free inpatient care.[14] The middle three groups we call the family-pay plans.

Other Covariates. In addition to dummy variables for each plan group, we also included covariates for age, sex, race, family income, health status, family size, and site (Manning et al., 1987). With the exception of family size and income, the data were collected before or at enrollment in the study. The value for family size varies by year. Family income data are from 1975 in Dayton, 1978 for the three-year group in South Carolina, and 1976 for all other participants.[15] Health status measures are described more fully in Brook et al. (1983, 1984), R. Burciaga Valdez et al. (1985), and Valdez (1986).

Although we have not tested for all possible interactions among covariates, we did examine some that are important for policy purposes (for example, income and plan). As a result, we have included interactions between being a child and plan in the inpatient and outpatient use equations (see below), between plan and income in the probabilities of any use of medical and of any inpatient use (see below), and between sex and age in all equations. The remaining interactions

[14] Differences among plans with 5, 10, and 15 percent upper limits are too small to detect at the level of annual expenditure. Hence, we have pooled across these different expenditure ceilings. See Keeler et al. (1987) for further discussion of how a varying ceiling affects demand.

[15] The first year of participation was 1975 for the Dayton participants; the South Carolina 3-year group began participation in late 1978 (about a quarter participated for two months and another quarter for one month of 1978); and the remainder of the sample enrolled in 1976 or early 1977. Most of the enrollment was in the latter half of 1976. We used these data because we believed the income measure was more reliable than the data on pre-experimental income. The data we used were collected on forms keyed to income tax returns, whereas data on pre-experimental income were responses to a personal interview.

were neither significant nor appreciable, and have been omitted.

C. *Unit of Analysis*

The unit of analysis is a person-year. We use the year as the time frame for ease of interpretation and because the upper limit on out-of-pocket expenses is an annual limit. We use the person as the unit of observation because most major determinants of the use of services are individual (for example, age, sex, and health status) rather than family (for example, insurance coverage, and family income).

II. Statistical Methods

We estimated two types of models. In addition to simple means (ANOVA), we present more robust estimates based on a four-equation model developed by Duan et al. (1982 and 1983). This model gains over ANOVA (and ANOCOVA) by exploiting three characteristics of the distribution of medical expenses. First, a large proportion of the participants use no medical services during the year. Second, the distribution of expenses among users is highly skewed. Third, the distribution of medical expenses is different for individuals with only outpatient use than for individuals with inpatient use.

Because of these three characteristics, ANOVA (and ANOCOVA) yields imprecise, though consistent, estimates of the effects of health insurance, health status, and socioeconomic status on the use of medical services, even for a sample size on the order of 20,000 (not all independent) observations. As Duan et al. (1982 and 1983) and Manning et al. (1987) show, a four-equation model that exploits the characteristics of the medical expense distribution yields consistent estimates with lower mean square error than ANOVA.

A. *The Four-Equation Model*

We partition the participants into three groups: nonusers, users of only outpatient services, and users of any inpatient services. We examine the expenses of the last two groups of users separately.

The first equation of the model is a probit equation for the probability that a person will receive any medical service during the year—from either inpatient or outpatient sources. Thus, this equation separates users from nonusers, and addresses the first characteristic described above, a large proportion of the population does not use medical services during the year. The second equation is a probit equation for the conditional probability that a user will have at least one inpatient stay, given that he has some medical use. This equation separates the two user groups, and thus addresses the third characteristic noted above, different distributions of medical expenses for inpatient and outpatient users.

The third equation is a linear regression for the logarithm of total annual medical expenses of the outpatient-only users. The fourth equation is a linear regression for the logarithm of total annual medical expenses for the users of any inpatient service. This last equation includes both outpatient and inpatient expenses for users of any inpatient services.[16]

The logarithmic transformation of annual expenses practically eliminates the undesirable skewness in the distribution of expenses among users, the second characteristic noted above. In particular, the logarithmic transform yields nearly symmetric and roughly normal error distributions. Further details are available in Duan et al. (1982 and 1983) and Manning et al. (1987).

While our use of the four-equation model is motivated by our desire to have the stochastic term approximate the normal assumption as closely as possible (to obtain robust estimates), the error distributions for the two levels of expense equations still deviate from the normal assumption. As a re-

[16]Grouping expenses by person rather than the more natural all-inpatient and all-outpatient expenditure eliminates the need to account for across-equation correlation in calculating standard errors of total expenditure.

sult, if we were to use the normal theory retransformation from the logarithmic scale to the raw dollar scale ($\exp(\sigma^2/2)$), the predictions would be inconsistent. Instead we use a nonparametric estimate of the retransformation factors, the smearing estimate, developed by Duan (1983), which in this application is the sample average of the exponentiated least squares residuals:

$$(1) \qquad \hat{\phi}_j = \sum_i \exp(\hat{\varepsilon}_{ij})/n_j, \quad j = 3,4,$$

where n_j = sample size for equation j,

$\hat{\varepsilon}_{ij} = \ln(y_{ij}) - x_{ij}\hat{\beta}_j$,

$\hat{\beta}_j$ = OLS estimate of β_j,

and i indexes the person. The smearing estimate is weakly consistent (asymptotically unbiased) for the retransformation factor if the error distribution does not depend on the characteristics x_i.[17]

A consistent estimate of the expected medical expense for person i based on the four-equation model is given by

$$(2) \quad E(\textit{Medical Expenditure}_i)$$
$$= \hat{p}_i\left[(1-\hat{\pi}_i)\exp(x_i\hat{\beta}_3)\hat{\phi}_3 + \hat{\pi}_i\exp(x_i\hat{\beta}_4)\hat{\phi}_4\right]$$

where $\hat{p}_i = \Phi(x_i\hat{\beta}_1)$ = estimated probability of any medical use,

$\hat{\pi}_i = \Phi(x_i\hat{\beta}_2)$ = estimated conditional probability for a medical user to have any inpatient use,

$\exp(x_i\hat{\beta}_3)\hat{\phi}_3$ = estimate of the conditional expense for medical services if outpatient only,

$\exp(x_i\hat{\beta}_4)\hat{\phi}_4$ = estimate of the conditional expense for medical services if any inpatient,

$\hat{\phi}_3, \hat{\phi}_4$ = estimated retransformation ("smearing") factors of the error terms for level of outpatient only and any inpatient expenditure equations.

Our estimates of predicted expenditure presented below are based on equation (2). We use equation (2) to predict medical expenditure for each person we enrolled, alternatively placing that person on each plan (by successively turning on plan dummy variables). We then average within plans over each predicted value to obtain a mean value for each plan. Standard errors of the predicted values are obtained by the delta method (see Duan et al., 1983, pp. 40, 48). The regression equations underlying our predicted values are presented in Manning et al. (1987).

B. *Correlation in the Error Terms*

Although we have over 20,000 observations, we do not have the same number of independent observations, because of substantial positive correlations in the error terms among family members and over time among observations on the same person. These correlations exist in all four equations. Failure to account for them in the analysis would yield inefficient estimates of the coefficients and inconsistent estimates of the standard errors. In the results presented below we have corrected the inference statistics (t, F, and χ^2) for this positive correlation using a nonparametric approach.[18]

C. *Selection Modes*

The econometric literature provides an additional class of models for continuous but limited dependent variables such as medical

[17] Moreover, when the normal assumption does hold, the smearing factor has high efficiency (90 percent or more) relative to the normal retransformation for a wide range of parameter values, including those in this analysis (see Duan, 1983, Section 5; and F. Mehran, 1973). In the results presented below, the smearing factors for the log level of expense for outpatient only users are estimated separately by plan and year to allow for heteroscedasticity. For the log level of expenses for users of any inpatient services, the smearing factor is a constant. See Duan et al. (1983) and Appendix C of Manning et al. (1987) for a comparison of normal theory and nonparametric retransformations.

[18] The correction is similar to that for the random effects least squares model, or equivalently the intracluster correlation model (S. R. Searle, 1971). The model is described in Brook et al. (1984), based on prior work by P. J. Huber (1967) on the variance of a robust regression.

expenditure. These models include the Tobit model (James Tobin, 1958), the Adjusted Tobit model (Wynand van de Ven and Bernard van Praag, 1981a,b), and sample selection models (G. S. Maddala, 1983). Like our four-equation model, these are multi-equation models, with an equation (often a probit) for whether there is a positive amount, and another equation for the level of the positive amount. These models differ from ours in that they explicitly model the correlation between the probability of any use and the level of use. Although they may appear to be more general, in fact for this problem they are not (Duan et al., 1984). In particular, the four-equation model just described is not nested within the sample selection model. Manning et al. (1987) provides a fuller discussion of these models and, using a split-sample validation, show that the four-equation model has significantly less bias than the sample selection model and is statistically indistinguishable on the basis of mean square error.[19] In a separate Monte Carlo study, Manning, Duan, and William Rogers (forthcoming) show that models such as the four-equation model can be more robust, and are no worse than selection models when the data are truly generated by a selection model.

III. Empirical Results

A. *Main Effects of Insurance Plan: ANOVA Estimates*

The data from the Health Insurance Experiment (HIE) clearly show that the use of medical services responds to changes in the amount paid out-of-pocket. Table 2 provides the sample means and standard errors by plan for several measures of use of services —the probability of being treated, visit and admission rates, and total expenses. The per capita expenses on the free plan (no out-of-pocket costs) are 45 percent higher than those on the plan with a 95 percent coinsurance rate, subject to an upper limit on out-of-pocket expenses. Spending rates on plans with an intermediate level of cost sharing lie between these two extremes. The right-most column shows that adjusting for the site imbalance in plan assignments (see Table 1) makes little difference.

Cost sharing affects primarily the number of medical contacts, rather than the intensity of each of those contacts. In other words, the differences in expenditures across plans reflect real variation in the number of contacts rather than an increase in the intensity or charge per service.[20] For example, outpatient expenses on the free plan are 67 percent higher than those on the 95 percent plan, while outpatient visit rates are 66 percent higher.

The largest decreases in the use of outpatient services occurs between the free and 25 percent plans, with smaller but statistically significant differences between the 25 percent and other family coinsurance (pay) plans ($\chi^2(2) = 9.48$, $p < .01$).

There are no significant differences among the family coinsurance (25, 50, and 95 percent) plans in the use of inpatient services. For the probability of any inpatient use, total admission rates, and inpatient expenses, the contrasts between the 25, 50, and 95 percent plans have p values greater than 0.50. As noted above, this lack of a significant difference is probably due to the effect of the upper limit on out-of-pocket expenses. Seventy percent of people with inpatient care exceeded their upper limit. Hence, the out-of-pocket cost of a hospitalization was at most \$1000 (in current dollars), and did not vary much among the pay plans (other than the individual deductible).[21]

[19] The bias in the selection models in the forecast sample was appreciable, on the order of 10–25 percent of the mean in the two replications we made ($p < .10$). In contrast, the bias for the 4-part model was 2 percent ($t = .50$).

[20] Keeler and John Rolph (1982) found that cost sharing affected the number of episodes of treatment, rather than the size of the episode. They used data from the first three years of the Dayton site. Kathleen Lohr et al. (1986) found a similar result for diagnosis-specific episodes.

[21] This is a good example of the difference between the response to a marginal price or coinsurance and the response to plan.

TABLE 2—SAMPLE MEANS FOR ANNUAL USE OF MEDICAL SERVICES PER CAPITA

Plan	Face-to-Face Visits	Outpatient Expenses (1984 $)	Admissions	Inpatient Dollars (1984 $)	Prob. Any Medical (%)	Prob. Any Inpatient (%)	Total Expenses (1984 $)	Adjusted Total Expenses (1984 $)[a]
Free	4.55	340	.128	409	86.8	10.3	749	750
	(.168)	(10.9)	(.0070)	(32.0)	(.817)	(.45)	(39)	(39)
25 Percent	3.33	260	.105	373	78.8	8.4	634	617
	(.190)	(14.70)	(.0090)	(43.1)	(1.38)	(0.61)	(53)	(49)
50 Percent	3.03	224	.092	450	77.2	7.2	674	573
	(.221)	(16.8)	(.0116)	(139)	(2.26)	(0.77)	(144)	(100)
95 Percent	2.73	203	.099	315	67.7	7.9	518	540
	(.177)	(12.0)	(.0078)	(36.7)	(1.76)	(0.55)	(44.8)	(47)
Individual Deductible	3.02	235	.115	373	72.3	9.6	608	630
	(.171)	(11.9)	(.0076)	(41.5)	(1.54)	(0.55)	(46)	(56)
Chi-Squared (4)[b]	68.8	85.3	11.7	4.1	144.7	19.5	15.9	17.0
P Value for *chi*-Squared (4)	<.0001	<.0001	.02	n.s.	<.0001	.0006	.003	.002

Note: All standard errors (shown in parentheses) are corrected for intertemporal and intrafamily correlations. Dollars are expressed in June 1984 dollars. Visits are face-to-face contacts with MD, DO, or other health providers; excludes visits for only radiology, anesthesiology or pathology services. Visits and expenses exclude dental care and outpatient psychotherapy.

[a]The figures in this column are adjusted for the imbalance of plans across sites as follows: the site-specific responses on each plan (simple means by site) are weighted by the fraction of the sample in each site and summed across sites. In the case of the 50 percent plan, which has no observations in Seattle, the weights are renormalized excluding Seattle.

[b]The *chi*-square statistic with 4 d.f. tests the null hypothesis of no difference among the five plan means. The *chi*-square statistic is a Wald test from the robust estimate of the information matrix (see Brook et al., 1984, for further details). It is used in lieu of the usual *F*-statistic because of the difficulty of computing such a statistic while allowing for intertemporal and interfamily correlation.

The Individual Deductible plan exhibits a somewhat different pattern from the other cost sharing plans. Recall that this plan has free inpatient care, but a 95 percent coinsurance rate (up to a $150 per person, or $450 per family annual maximum) for outpatient services. Total expenditures on this plan are significantly less than the free plan ($t = -2.34$, $p < .02$). This overall response is the sum of a one-third reduction in outpatient expenses ($t = -6.67$), and a less than one-tenth reduction in inpatient expenses ($t = -0.68$). Thus, this plan looks like a combination of the 50 or 95 percent plans for outpatient care and the free of 25 percent plan for inpatient care. The admission rate for the Individual Deductible plan lies roughly midway between the free plan and family coinsurance plan rates, suggesting a nontrivial cross-price elasticity between inpatient and outpatient services.

B. *Main Effects of Insurance Plan: Four-Equation Estimates*

Because sample means are quite sensitive to the presence of catastrophic cases, we used the four-equation model to provide more robust estimates of the plan responses.[22] The use of covariates in these equations further enhances precision and re-

[22]For example, the ANOVA estimates of the response to cost sharing for total expenses (not adjusted for site) show a statistically insignificant reversal between the 50 and 25 percent plans. Although such a reversal is compatible with theory (due to the MDE) the reversal is almost certainly due to chance. One participant on the 50 percent plan had a very expensive hospitalization (total medical expenses of $148,000 in one year); that single observation, which was the largest observation in the entire sample, adds $106 dollars to the 50 percent plan mean (16 percent of that plan's mean).

TABLE 3—VARIOUS MEASURES OF PREDICTED MEAN ANNUAL USE OF MEDICAL SERVICES, BY PLAN

Plan	Likelihood of Any Use (%)	One or More Admissions (%)	Medical Expenses (1984 $)
Free	86.7	10.37	777
	(0.67)	(0.420)	(32.8)
Family Pay			
25 Percent	78.8	8.83	630
	(0.99)	(0.379)	(29.0)
50 Percent	74.3	8.31	583
	(1.86)	(0.400)	(32.6)
95 Percent	68.0	7.75	534
	(1.48)	(0.354)	(27.4)
Individual Deductible	72.6	9.52	623
	(1.14)	(0.529)	(34.6)

Note: Standard errors are shown in parentheses. Medical services exclude dental and outpatient psychotherapy. The predictions are for the enrollment population carried forward through each year of the study. The standard errors are corrected for intertemporal and intrafamily correlation. The t-statistics for the contrasts with the free plan are -6.69, -6.33, -11.57, and -10.69 for the last four rows of the first col., respectively; -2.74, -3.57, -4.80, and -1.28 for the last four rows of the second col., respectively, and -4.05, -4.91, -6.74, and -3.78 for the last four rows of the third col., respectively. These t-statistics are larger than those one would compute from the standard errors shown in the table because use of the standard errors ignores the positive covariance between the two predicted plan means from the shared $X\beta$ terms. The differences in expenses between the 25 and 50 percent plans are significant at the 5 percent level ($t = 1.97$), and between the 50 and 95 percent plans are significant at the 6 percent level ($t = 1.93$). The parameter estimates underlying these predictions are available in Manning et al. (1987).

moves the relatively minor imbalances across plan, including the site imbalance. Table 3 presents estimates from this model of plan response for the probability of any use of medical services, the unconditional probability of any inpatient use, and total medical expenses. Figure 1 displays the expenditure results.

Mean predicted expenditure in the free care plan is 46 percent higher than in the 95 percent plan ($p < .001$), almost exactly the difference found in the sample means.[23] Like

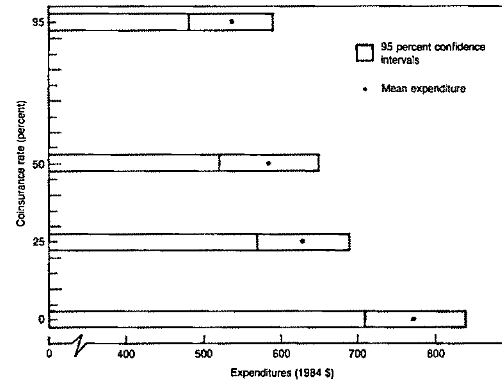


FIGURE 1. DEMAND AND 95 PERCENT CONFIDENCE INTERVALS BY COINSURANCE RATE

the sample means, these more robust estimates also indicate that the largest response to plan occurs between free care and the 25 percent plan, with smaller decreases thereafter.

Not surprisingly, given the approximate orthogonality of plan and covariates, adding covariates does not change the estimated probability of any use of medical services—87 percent of the free plan participants are predicted to use any service during the course of the year, while only 68 percent of the 95 percent plan participants are. These differences in the likelihood of receiving any care account for over three-fifths of the overall response to cost sharing. Virtually all the remaining response is attributable to the effect of cost sharing on hospital admissions.

Cost sharing for outpatient services only (the individual deductible plan) produces a different pattern of utilization than cost sharing for all services. Outpatient-only cost sharing reduces total expenditures relative to free care ($p < .0001$), largely by reducing the likelihood of any use ($p < .0001$). Outpatient-only cost sharing also reduces inpatient use, but by an insignificant amount ($p = .20$ for the probability of any inpatient use). This last result is the only important change

[23] It may seem that this is a trivial result that follows from the orthogonality of plan and covariates. Such is not the case because of the nonlinear transformations in the 4-part model. Using the logarithm of expenditure plus $5, for example, as a dependent variable instead of the 4-part model would lead to a much larger estimate of plan response, one that would be biased upward. (See Duan et al., 1983; Manning et al., 1987.)

TABLE 4—VARIOUS MEASURES OF PREDICTED ANNUAL USE OF MEDICAL SERVICES, BY INCOME GROUP

Plan	Income: Lowest Third Mean	Income: Middle Third Mean	Income: Highest Third Mean	Significance Tests t on Contrast of: Middle vs. Lowest Thirds[a]	Significance Tests t on Contrast of: Highest vs. Lowest Thirds[a]
Likelihood of Any Use (Percent)					
Free	82.8	87.4	90.1	4.91	5.90
Family Pay					
25 Percent	71.8	80.1	84.8	5.45	6.28
50 Percent	64.7	76.2	82.3	4.35	4.86
95 Percent	61.7	68.9	73.8	3.96	4.64
Individual Deductible	65.3	73.9	79.1	6.09	7.09
Likelihood of One or More Admissions (Percent)					
Free	10.63	10.14	10.35	−0.91	−0.35
Family Pay					
25 Percent	10.03	8.44	7.97	−2.95	−2.75
50 Percent	9.08	8.06	7.77	−1.78	−1.66
95 Percent	8.77	7.38	7.07	−2.79	−2.46
Individual Deductible	9.26	9.44	9.88	0.31	0.68
Expenses (1984 $)					
Free	788	736	809	−1.78	0.53
Family Pay					
25 Percent	680	588	623	−3.17	−1.47
50 Percent	610	550	590	−1.89	−0.49
95 Percent	581	494	527	−3.09	−1.41
Individual Deductible	609	594	670	−0.57	1.38

Note: Excludes dental and outpatient psychotherapy. Predictions for enrollment population carried forward for all years of the study.

[a]The t-statistics are corrected for intertemporal and intrafamily correlation. The statistics test the null hypothesis that the mean of middle (highest) third equals the mean of the lowest third; for example, the 4.91 figure implies we can reject at the .001 level the hypothesis that in the free plan the likelihood of any use for the lowest and middle thirds of the income distribution are equal.

from the previously published analysis of the first 40 percent of the data (Newhouse et al., 1981). In that analysis, inpatient use was less on the deductible plan, and one could reject at the 5 percent level the hypothesis that the free plan and individual deductible plan means for inpatient use were the same. This difference may have occurred because inflation in the late 1970's reduced the real value of the deductible, which was kept fixed at $150 (i.e., in nominal dollars), or may have simply been due to chance.

C. *Use by Subgroups*

An important goal of the HIE was to study how the response to cost sharing varied across subgroups. These included differences in responses across income groups, differences between adults and children, differences between the sickly and healthy, as well as differences across time (for example, any transitory surges in use as insurance changed), and differences across medical markets (for example, urban vs. rural).

1. *Across Income Groups.* Different aspects of the use of medical services exhibit different responses to income (Table 4).[24] In Ta-

[24]Recall that the income measure comes from the first partial year of enrollment.) The division into thirds is site specific (for example, the lowest third is the lowest third of each site's income distribution), because

ble 4 we observe differences in use that are due to both income directly and the effects of variables correlated with income; that is, these are not partial effects.

Within each of the five plans the probability of any use of medical services increases with income, with larger increases for the family pay (25, 50, and 95 percent) and individual deductible plans than the free plan.[25] In contrast, the (unconditional) probability of any use of inpatient services declines with income for the family pay plans, but is not significantly different across income groups for the two plans with free inpatient care (the free and individual deductible plans). Because of these two conflicting effects of income—positive on outpatient use but negative on inpatient use—the net result on total expenditure is a shallow U-shaped response.

Our estimate of the differences by income group within the family-pay plans is influenced by the income-related upper limit in out-of-pocket expenses. The observed response in a combination of the direct response to income, and the fact that families with lower incomes are more likely to exceed their (lower) limit and receive free care for part of the year.[26] If medical care is a normal good, then any positive direct effect of greater income would be reduced by the decreased likelihood of going over the limit. In the case of the positive effect of income on the probability of any use, the direct income effect is probably more important, and in the case of the negative effect on the probability of any inpatient use, the limit has relatively more influence.[27]

The Individual Deductible plan provides a cleaner test of the differences by income group of use of medical services, because the deductible in that plan is not income related. We observe an insignificant 10 percent increase in medical expenses between the bottom and top third of the income distribution. The effect of income is limited to an increased likelihood of using outpatient services, probably because inpatient services are free on this plan.

Thus far we have compared response among income groups rather than examining the partial effect of income. Although income has a statistically significant positive partial effect on use of service, the magnitude is small enough to be swamped by other factors correlated with income (for details see Manning et al., 1987, Appendix A, Tables 2–4 and 6).[28]

2. *Across Age Groups*. We found about the same outpatient response to insurance plans for children (ages less than 18) as for adults, but children are less plan responsive for inpatient care (Table 5).[29] As we observed with a subset of these data (see Newhouse et al., 1981 and 1982; Leibowitz et al., 1985), we cannot reject the hypothesis that admission rates for children show no response to insurance coverage.[30] By contrast, adults

1) expenses are not corrected for cross-sectional differences in prices, and 2) we did not want to confound income and site; the sites were chosen to represent a spectrum of medical market characteristics. See Manning et al. (1987, Table 1, Appendix D) for the ANOVA estimates by plan income group (as well as by other subgroups).

[25] Note that this is not a *ceteris paribus* statement, so there is no contradiction with standard theory, which would suggest no income effect in the free plan.

[26] See Manning et al. (1987, Appendix B) for data on the proportion exceeding the upper limit on out-of-pocket expenses.

[27] Some may argue that income is endogenous with respect to inpatient expenditure. This may well be true, but is not likely to account for our result because only a few months of data are "tainted."

[28] Income has a moderately significant (at $p < .10$) and positive partial effect on use in all but the inpatient expenditure equation; in the level of outpatient-only expenditures, however, the income coefficients are of mixed sign. The probabilities with which we can reject the null hypothesis that the income coefficients are zero are: $p < .001$ for any use of medical services, $p < .10$ for the probability of any inpatient use given any medical use, $p < .001$ for the (log) level of outpatient-only use, and $p > .10$ for (log) level of medical expenditure if any inpatient use. The test statistics include plan income interactions and missing value replacement dummy variables.

[29] Recall that children are overrepresented in the study relative to the population of our sites. Hence, our estimates understate (modestly) the population responsiveness in our sites.

[30] $\chi^2(4) = 5.19$ using ANOVA estimates for the probability of any inpatient use, and $\chi^2(4) = 5.36$ for the admission rate. Another possible hypothesis is no differential plan response for children relative to adults.

TABLE 5—VARIOUS MEASURES OF PREDICTED ANNUAL USE OF MEDICAL SERVICES, BY AGE GROUP AND PLAN

Plan	Likelihood of Any Use (%) Mean	One or More Admissions (%) Mean	Medical Expenses (1984 $) Mean
Children			
Free	84.0	5.33	346
Family Pay			
25 Percent	75.1	4.98	287
50 Percent	70.3	4.62	279
95 Percent	63.5	4.23	236
Individual Deductible	68.5	5.86	299
Adults			
Free	88.6	13.9	1080
Family Pay			
25 Percent	81.4	11.5	872
50 Percent	77.1	10.9	797
95 Percent	71.2	10.2	744
Individual Deductible	75.6	12.1	852

Note: Excludes dental and outpatient psychotherapy services. The eight t-statistics for the contrasts between the free plan and the pay plans for the likelihood of any use all exceed 6. For one or more admissions, the t-statistics for children for contrasts with the free plan (rows 2–5) are 0.55, 1.13, 1.81, and -0.63, respectively, and for adults are 2.92, 3.64, 4.69, and 1.89, respectively (for example, the t-statistic on the difference between 13.9 and 12.1 is 1.89). For medical expenses the t-statistics on contrasts with the free plan for children are 2.16, 2.20, 4.10, and 1.42, respectively, and for adults are 3.70, 4.80, 6.07, and 3.63, respectively.

have significantly lower use of inpatient services on the family-pay plans than they do on the free plan.[31] For outpatient services, we observe a very similar pattern of plan responses for children and adults.

D. *Other Subgroups*

1. *Health Status.* Although health status was a strong predictor of expenditure levels, we observed no differential response to health insurance coverage between the healthy and the sickly (Manning et al., 1987). This null result is striking because of the upper limit feature. If anything, the presence of an upper limit on out-of-pocket expenses would lead to less plan response for the sickly; all other things equal, sicker individuals are more likely to exceed their upper limit and receive some free care—especially on the 95 percent plan, where care is free after gross expenditures of $1050 or more. Furthermore, some might expect the sickly to be less responsive to insurance coverage than the healthy, on the supposition that their use of services is less discretionary. If, in fact, there is no interaction between plan and health status, one can infer that the opposite is true at the margin; that is, at the margin the sickly exhibit more discretion.

2. *Sites.* The six sites in the HIE were selected to reflect a spectrum of city sizes, waiting times to appointment, and physician to

We can reject this hypothesis; the test statistics are $\chi^2(4) = 16.49$ for the probability of any inpatient use and $\chi^2(4) = 14.08$ for total admissions. Hence, it appears that children and adults respond differently and that children do not respond to cost sharing for inpatient care.

[31] $\chi^2(3) = 24.22$ for the probability of any inpatient use and 16.31 for the admission rate. By contrast, there are no significant differences among the family pay plans for adults. $\chi^2(2) = 1.69$ for expenditures, 0.73 for total admissions, and 1.39 for the probability of any inpatient use, again based on ANOVA (see Manning et al., 1987, Table 2, Appendix D for the ANOVA estimates).

P4157

TABLE 6—USE OF DENTAL SERVICES BY DENTAL PLAN: SAMPLE MEANS

Dental Insurance Plan	Year 1 of Dental Coverage: Probability (%)	Year 1: Visits	Year 1: Expenses per Enrollee ($)	Year 2 of Dental Coverage: Probability (%)	Year 2: Visits	Year 2: Expenses per Enrollee ($)
Free	68.7	2.50	380	66.8	1.93	261
	(1.19)	(.065)	(18.0)	(1.18)	(.049)	(12.5)
25 Percent	53.6	1.73	224	52.6	1.51	190
	(3.39)	(.138)	(32.8)	(3.34)	(.111)	(28.0)
50 Percent	54.1	1.80	219	53.0	1.50	177
	(2.41)	(.118)	(31.3)	(2.55)	(.103)	(32.3)
95 Percent	47.1	1.39	147	48.3	1.44	179
	(2.59)	(.098)	(18.7)	(2.62)	(.099)	(24.9)
Individual Deductible	48.9	1.70	242	48.1	1.33	158
	(2.12)	(.104)	(24.1)	(2.12)	(0.080)	(20.4)

Note: Expenses were converted to January 1984 dollars using the dental fee component of the Consumer Price Index. There has been no adjustment for regional differences in prices, or differences in population characteristics across plans and years. Standard errors (shown in parentheses) are corrected for intrafamily correlations.

population ratios (Newhouse, 1974).[32] Our concern was that the response to insurance coverage could vary according to the complexity of the medical market or to the excess demand in the medical delivery system. Yet we found no differences among the sites in the response to insurance coverage, $\chi^2(19) = 14.96$ ($p > .50$). The uniformity of response across the sites gives some reason to believe the results may be representative of the United States, and we have so used them below.

Interestingly, the site with the longest delay to appointment and lowest physician to population ratio (Fitchburg) had the second highest probability of any use, the second highest expenditures per enrollee, and the highest probability of any inpatient use. The latter two phenomena may represent substitution of inpatient for outpatient care (Jeffrey McCombs, 1984), and the first may indicate that the presence of emergency rooms removes the constraint of the queue (Stephen Long, Russell Settle, and Bruce Stuart, 1986).[33]

2. *Period of Enrollment.* As noted above, we enrolled families for three or five years to see if the response to insurance changed over time and if the duration of enrollment mattered. The free plan might generate transitorily high demand; the 95 percent plan might generate postponement of demand at the end of the experiment (Arrow, 1975; Metcalf, 1973). Neither effect was found; see Manning et al. (1987) for further details.[34] Nor did duration of enrollment matter to either the absolute level of spending or the responsiveness to plan.

3. *Subexperiments.* As described above, the Health Insurance Experiment contained a number of subexperiments to study methods effects. None of the subexperiments had a measurable effect on expenditure (Manning et al., 1987).

[32]For example, city sizes in 1970 ranged from 34,000 (Georgetown County) to 1.2 million (Seattle), waiting times for nonemergent care in 1973–74 ranged from 4.1 days (Seattle) to 25.0 days (Fitchburg), and physicians per capita in 1972 ranged from 30 per 100,000 (Fitchburg) to 59 per 100,000 (Seattle).

[33]Length of waiting time to an appointment with a primary care physician is associated positively with the use of emergency rooms (O'Grady et al.).

[34]A transitory effect was found for dental services; see Manning et al. (1985, 1986a) for details.

E. *Dental Results*

These results are reported in greater detail elsewhere (Manning et al., 1985, 1986a). Dental services do show greater responsiveness to plan in Year 1 than in subsequent years ($p < .001$) (Table 6). This would be expected if dental services were more durable than other medical services, as is plausible. The responsiveness of demand by plan in Year 2, which is typical of the middle years, is of the same general magnitude as that for other medical services.

F. *Health Status Outcome Results*

These results also are reported in greater detail elsewhere (Brook et al., 1983, 1984; Valdez et al.; Valdez; Howard Bailit et al., 1985). For the person with mean characteristics, we can rule out clinically significant benefits from the additional services in the free fee-for-service plan relative to either the cost-sharing plans or the HMO experimental group. For poor adults (the lowest 20 percent of the income distribution) who began the experiment with high blood pressure (specifically, who were in the upper 20 percent of the diastolic blood pressure distribution) there was a clinically significant reduction in blood pressure in the free fee-for-service plan compared to the plans with cost sharing. Epidemiologic data imply that the magnitude of this reduction would lower mortality about 10 percent each year among this group, about 6 percent of the population. (The sample size is much too small to test this prediction with actual mortality among the experimental population.) For poor adults who began the experiment with vision problems that were correctable with eyeglasses, there was a modest improvement in corrected vision. Individuals on the free care plan between the ages of 12 and 35 showed a modest improvement in the health of the gums; caries (decayed teeth) were also more likely to be filled on the free care plan.

The specific gains in health just described, for high blood pressure, myopia, and dental care, were all for relatively prevalent chronic problems (of course, we had difficulty detecting effects for rare problems) that are relatively inexpensive to diagnose and remedy. One can infer that programs targeted at these problems would be much more cost effective in achieving these gains in health than free care for all services. For example, more than half the benefit of free care for high blood pressure (and presumably for risk of dying) was available from a one-time screening examination, whose cost is a small fraction of free care for all services (Keeler et al., 1985).

G. *Health Maintenance Organization Results*

We also randomized a group of participants into an HMO, the Group Health Cooperative of Puget Sound in Seattle.[35] This group, whom we call the HMO Experimentals, was given a plan of benefits identical to the free fee-for-service (FFS) plan. In addition, we enrolled a random sample of existing HMO enrollees, the HMO Controls. Thus, a comparison of the experimentals and the free fee-for-service plan establishes the "pure" HMO effect on use; a comparison of the experimentals and controls establishes the extent, if any, of selection with respect to the HMO.[36]

Our results (Table 7) show no evidence of selection in the single HMO that we studied; those previously enrolled at the HMO (the Controls) used services at approximately the same rate as those who were not previously enrolled (the Experimentals). By contrast, the percentage of Experimental plan participants with one or more hospital admissions was only two-thirds as great as the percentage on the free fee-for-service plan. Because outpatient use was approximately similar on the two plans, the expenditure difference between the HMO Experimentals and free fee-for-service participants was

[35]An HMO is reimbursed a fixed amount per month, in return for which it agrees to provide medical care. Thus, unlike fee-for-service medicine, the approximate marginal revenue from delivering additional services is zero. Of course, there are market constraints on the HMO's behavior because is competes with fee-for-service medicine for patients.

[36]The fee-for-service sample in this comparison is from Seattle, in order to keep the population sampled the same between the two groups.

TABLE 7—ANNUAL USE OF MEDICAL SERVICES PER CAPITA, SEATTLE SAMPLE, BY HMO AND FFS STATUS[a]

Plan	Likelihood of Any Use (%)	One or More Admissions (%)	Imputed Expenditures ANOVA[b] (1983 $)	Imputed Expenditures with Age-Sex Covariates[b] (1983 $)	Person Years
HMO Experimental	87.0	7.1	434	426	3687
	(1.0)	(0.50)	(28)	(23)	
HMO Control	91.1	6.4	432	465	2596
	(0.8)	(0.55)	(34)	(47)	
Free Fee-for-Service	85.3	11.2	640	612	1221
	(1.6)	(1.17)	(81)	(66)	
t-Statistic on Free-Experimental Difference[c]	−0.88	3.24	2.44	2.69	
p Value for t-Statistic, 2 tail	n.s.	.0012	.016	.007	

[a]Standard errors are shown in parentheses. The sample includes participants while they remained in the Seattle area. The sample excludes children born into the study and excludes partial years except for deaths, similar to Tables 1 and 2 above. For HMO Controls and Experimentals, the data include both in- and out-of-plan use. The standard errors are corrected for intertemporal and intrafamily correlation using an approach due to Huber in a similar fashion to Tables 1 and 2 above. The numbers differ slightly from those in Manning et al. (1984), because of minor corrections in the data, as well as the use of a less precise, but more robust method of calculating standard errors. The method is the same as that described in Table 2.

[b]See Manning et al. (1984) for details of imputation method.

[c]Testing null hypothesis of no difference between HMO Experimental and Free Fee-for-Service plan.

somewhat narrower; expenditures per person among the HMO Experimentals were only 72 percent of expenditures on the free fee-for-service plan.

These findings demonstrate that a markedly less hospital-intensive style of medicine than is commonly practiced in the fee-for-service system is technically feasible. Whether the technical style will be attractive to consumers, and, if it is, whether a market of competing HMOs is economically feasible —or whether adverse selection problems will prove insurmountable (Michael Rothschild and Joseph Stiglitz, 1976)—are still somewhat open questions, although the size and history of large HMOs such as Group Health Cooperative of Puget Sound suggest that the style is attractive to some consumers.

In projecting the effect of the growing HMO market share on hospital admissions and medical expenditure, one must keep in mind that the above comparisons have been made against the free care plan. Because virtually all private fee-for-service health insurance plans include some cost sharing, one should compare the reduction in hospital admissions at the HMO, some 35 percent, with the reduction caused by cost sharing, some 15 to 25 percent depending on plan. The values presented above, however, do represent the *ceteris paribus* HMO effect; if an HMO were to use cost sharing, its observed rates of use might be even lower.

Consumers contemplating enrollment in an HMO will weigh the cost savings against any effect of the reduction in services upon health status and consumer satisfaction. Our findings on health status of the HMO are analogous to those in the free fee-for-service system; the mean person in the fee-for-service plan appeared to derive few or no benefits from the additional hospital services (Ware et al., 1986; Elizabeth Sloss et al. 1987). Those who are both in poor health and of low income who were in the HMO exhibited a higher rate of bed-days and serious symptoms (relative to those in the free fee-for-service plan). There is thus some suggestive evidence that special programs to facilitate access for Medicaid enrollees in HMOs may be worthwhile, but we caution that this result comes only from one HMO

(albeit a well-established and well-regarded HMO) and that the precision with which we could measure results among the poor, sick group makes this result less than definitive, even in the case of this HMO.

Those who had self-selected the HMO (the Controls) were on average as satisfied with their care as those in the fee-for-service system (Allyson Davies et al., 1986). Theory would suggest the marginal person would be equally satisfied in both systems, and it is not surprising that we detected no difference for the average person. By contrast, the HMO Experimentals were less satisfied overall with their care than those in the fee-for-service system, although on certain dimensions they were as satisfied or even more satisfied.

IV. Conclusions

A. *On Comparing our Estimates of Demand with those in the Literature*

Our results leave little doubt that demand elasticities for medical care are nonzero and indeed that the response to cost sharing is nontrivial. How do our estimates compare with those in the nonexperimental literature?

This question is difficult to answer, because most prior empirical work has parameterized cost sharing as a constant coinsurance rate (for example, Feldstein, 1971, 1977) or has examined particular changes in insurance plans (for example, an imposition of a $3 per visit copayment: Scitovsky and Snyder; Phelps and Newhouse, 1972; Scitovsky and McCall). By contrast, experimental policies were from a two-parameter family (coinsurance rate and maximum dollar expenditure). We make no apologies for this intentional noncomparability; a constant coinsurance rate, while convenient for obtaining comparative statics results, is not an insurance policy that theory suggests would be optimal, assuming risk aversion (Arrow, 1963, 1971, 1973, 1975). Indeed, an optimal policy would almost certainly contain a stop-loss feature, exactly as the experimental plans did.[37]

One could, of course, attempt to estimate the functional response of demand to variation in the two parameters; one can view the values presented above as selected points in the response surface generated by varying coinsurance at given maximum dollar expenditure levels. In order to compare our results with those in the literature, however, we must extrapolate to another part of the response surface, namely, the response to coinsurance variation when there is no maximum dollar expenditure. Although any such extrapolation is hazardous (and of little practical relevance given the considerable departure from optimality of such an insurance policy), we have undertaken such an extrapolation rather than forego entirely any comparison with the literature. Specifically, we have used three different methods to estimate a price elasticity comparable to the estimates in the literature:

1) One can estimate a pure coinsurance elasticity by analyzing variation in the demand for episodes of care rather than annual expenditure per person (Keeler and John Rolph, 1982; Keeler et al., 1987). The theory of demand suggests that individuals who have not yet exceeded the upper limit on out-of-pocket expenses, when making a marginal medical consumption decision, will discount the nominal price by the probability of exceeding the limit (because with that probability the true price is zero) (Keeler, Newhouse, and Phelps, 1977; Randall Ellis, 1986).[38] We therefore examine demand for episodes of treatment by individuals who are more than $400 from their limit. This gives an approximation of the pure price effect if such people treat the true probability of exceeding their limit as nearly zero.[39] The

[37]A stop-loss feature means there is a maximum out-of-pocket loss that the insured can sustain. In addition to its risk-reduction properties, no worst-case payment would have been possible without a stop-loss feature, and hence selection effects might have been introduced into the experiment.

[38]The specific result requires risk neutrality and separability of the utility function in health and money, but the qualitative results does not.

[39]Because there was no appreciable difference between demand for outpatient episodes when the MDE remaining was between $1 and $400 and when it was more than $400, this assumption seems reasonable for

TABLE 8—ARC ELASTICITIES FOR VARIOUS TYPES OF CARE CALCULATED FROM EPISODES[a]

Range of Nominal Coinsurance Variation	Type of Care					
	Outpatient[b]					
	Acute	Chronic	Well	All[c]	Hospital	All Care[c]
0–25 Percent	.16	.20	.14	.17	.17	.17
	(.02)	(.04)	(.02)	(.02)	(.04)	(.02)
25–95 Percent	.32	.23	.43	.31	.14	.22
	(.05)	(.07)	(.05)	(.04)	(.10)	(.06)

[a]The method of calculating standard errors (shown in parentheses) is described in Keeler et al. (1987).

[b]Acute conditions are unforeseen and treatment opportunities are nondeferrable. Chronic episodes comprise foreseen and continuing expenditure; treatment is designed to ameliorate the consequences of the disease rather than cure. Flare-up of chronic conditions, which are unforeseen, we treat as acute. Well care episodes are medically deferrable without great loss and can occur when the patient is not considered sick.

[c]Estimate derived by weighting elasticities for various types of care by budget shares.

estimation method controls for unobserved propensities to have episodes, as well as other observed covariates by looking at experience before and after the MDE is exceeded; see Keeler and Rolph for a description of the methodology. We have computed arc elasticities for the 0–25 and 25–95 percent ranges of coinsurance; those elasticities are shown in Table 8.

2) A second estimate comes from using an indirect utility function and applying it to total expenditure in the 25–95 percent range. This estimate is very close to the first, -0.18 (Manning, 1986).

3) A third estimate comes from a similar calculation to those in the literature, that is, it uses average coinsurance rates (Table 9). The usual proof of an upward bias in the elasticity estimate from using the average coinsurance rate (Newhouse, Phelps, and Marquis) does not apply here because of the balance across plans. The amount of bias, if any, depends on two effects that work in opposite directions. For small expenditures the experimental plans will exhibit smaller expenditure than would a pure coinsurance rate plan of 16 or 31 percent (because the effective coinsurance rate is likely to be higher); for large expenditures exceeding the MDE the opposite will be true (because the marginal coinsurance rate will be zero, not positive). Which effect predominates is an empirical question the experimental data cannot resolve; empirically, this method yields values that are somewhat lower but still close to those of the other two methods. (The lower value suggests the first bias predominates.)

In sum, these three methods suggest that price elasticities for a constant coinsurance policy are in the -0.1 to -0.2 range, values that are consistent with those in the lower range of the nonexperimental literature.

TABLE 9—ARC ELASTICITIES FOR VARIOUS TYPES OF CARE CALCULATED FROM AVERAGE COINSURANCE RATES

Range of Nominal Coinsurance Variation	Range of Average Coinsurance Variation	All Care	Outpatient Care
0–25 Percent	0–16	.10	.13
25–95 Percent	16–31	.14	.21

Source: Calculated from data in Table 2 (outpatient) and Table 3 (total). For those who wish to calculate arc elasticities with the 50 percent plan, from the data in Tables 2 or 3, the average coinsurance rate in the 50 percent plan is 24 percent.

outpatient episodes. It may cause some bias in the estimated hospital elasticity; if the true MDE were, say $10,000 rather than $1000, we might observe fewer hospitalizations.

B. *On the Explanation of the Sustained Rise in Medical Expenditure*

At first blush, our estimates of demand response imply that the spread of health insurance can account for only a modest portion of the postwar rise in medical expenditure, contrary to the commonly held view described in the introduction. Between 1950 and 1984, real medical expenditure rose by a factor of 7,[40] but our estimates of insurance elasticity do not begin to imply this degree of increase. To demonstrate this point, we use the average coinsurance rate. Despite its imperfect measure of the generosity of insurance, it is a gross measure of how much insurance changed over the post-1950 period and therefore indicative of the role insurance might have played in this increase. Table 10 shows the average coinsurance rate by type of service (see Table 9 for comparable values from the 25, 50, and 95 percent plans). Although the figures by service are based on an arbitrary accounting convention, they suggest that the change in insurance in the postwar period was of roughly the same absolute magnitude as the difference between the 95 percent coinsurance and free care plans.[41]

Because the free plan demand was only around 1.5 times that of the 95 percent plan, it appears that the change in insurance can explain only a small part, perhaps a tenth, of the factor of 7 change in health expenditure in the post-World War II period.

Nor can changes in real income (around a factor of 3 during this period) directly account for much of the rise. Income elasticities estimated from the experimental data (the partial response, not the one shown in Table 5) are at most 0.2—much too small to account for anything like a factor of 7 change.[42]

TABLE 10—CHANGE IN AVERAGE COINSURANCE RATE, 1950–84, BY TYPE OF SERVICE

Year	Hospital	Physician	Other	Total
1950	.30	.83	.86	.66
1984	.09	.28	.56	.28

Source: Levit et al. (1985).

Thus, we still must account for the bulk of the expenditure increase. The rather obvious "accounting" explanation of the expenditure increase is technological change; there are a host of new medical products and procedures today that did not exist in 1950. For example, those with kidney failure are now treated with renal dialysis or kidney transplantation; in 1950 these individuals died rather quickly. This merely pushes the puzzle back one stage, however; what role, if any, did insurance (and income growth) play in inducing the technological change? Unfortunately that question cannot be answered from experimental data.[43]

Thus, if insurance is playing a role in inducing a welfare loss, given the rate of increase in medical expenditure, the bulk of that loss must come from its having induced innovation for which unsubsidized consumers would not be willing to pay.[44] Given that most countries in the world have also experienced a long-term sustained increase in expenditure despite widely varying institutional arrangements, it is at least arguable that consumers would be willing to pay for much of the increase, but there clearly

[40] Nominal expenditure data from Katherine Levit et al. (1985) deflated by the GNP deflator.

[41] The accounting convention used by the Health Care Financing Administration allocates a common deductible to services in proportion to gross expenditure. We have followed the same convention in calculating comparable figures from the experimental data.

[42] Real GNP increased between 1950 and 1983 by a factor of 2.9. Even allowing for the usual downward bias from using measured income to estimate income elasticities, it is clear that changes in income can only explain a modest portion of the expenditure increase.

[43] Because most consumers have been insured for inpatient services throughout the relevant time period, it is an extremely difficult question to answer from nonexperimental data. Moreover, one does not observe insurance policies that do and do not cover new procedures, so there is no straightforward test of willingness to pay for new technology. Although virtually all policies do not cover "experimental" procedures, once efficacy and "safety" are demonstrated, insurance plans tend to cover all procedures.

[44] The willingness-to-pay calculation should include any willingness to pay for others' care.

has been no pure market test (Newhouse, 1977, 1984).

C. *On the Magnitude of Welfare Loss from Health Insurance*

Setting aside the issue of possible welfare loss from induced technological change, one can estimate the welfare loss in the usual static framework. Under a number of strong assumptions (including that gross medical care prices are competitive and there are no externalities), our estimates imply a nontrivial welfare loss from first-dollar health insurance coverage. An approximation to the loss from moving from a universal 95 percent plan (with a $1000 MDE) to the free care plan is $37 to $60 billion, as against an expenditure around $200 billion on these services in 1984 by the under 65 population.[45] From the $37–60 billion figure must be deducted some amount for the reduced risk in the free plan relative to the 95 percent plan. Usual values for risk aversion, however, would suggest the deduction is small in the presence of a $1000 cap (Feldstein, 1973; Keeler, Morrow, and Newhouse, 1977). Although the $37–60 billion figure is probably overstated by ignoring externalities and assuming medical care prices are competitive, it ignores any welfare loss from induced technological change.[46]

D. *On the Existing Insurance Coverage of Various Medical Services*

One can find several economic reasons for the traditionally more generous coverage of inpatient services relative to outpatient services (Table 10). Loading charges (as a per-

[45]The $37 and $60 billion figures are calculated in the usual Harberger fashion by taking the $325 per capita difference in spending between the 95 percent and free plans from Tables 3 and 6 (Year 2 values) and adding $19 for mental health services (Wells et al., inflated by the change in the CPI Medical Services prices index between 1977 and 1984). We then multiply by 207 million, the number of resident civilians under 65. This yields a figure of $71 billion. One then multiplies by 0.525 or 0.845. Both fractions are larger than the usual 0.5 because we do not start at an unsubsidized point. Our 95 percent $1000 MDE plan had an average coinsurance rate of 0.31. An upper bound on the welfare loss comes from assuming that individuals valued the last dollar at 0.31. A lower bound on the welfare loss comes from assuming that the extra spending is all from individuals who valued the last dollar of spending at 0.95, the nominal coinsurance rate. The 0.525 figure equals $1-.95/2$, and the 0.845 figure equals $1-.31/2$.

The $200 billion figure can be estimated in two ways: 1) Data from Levit et al. show expenditure on personal health care services of $342 billion in 1984. Waldo and Lazenby (1984, Table 11) estimate that $120 billion of this is for the over 65, leaving $222 billion for the under 65. Some of this, however, is for noncovered services, such as nonprescription drugs, and some other part is for ineligible populations, such as the institutionalized. Adjusting for these noncomparabilities is necessarily somewhat imprecise, but would probably leave a final figure around $200 billion. 2) Data from Tables 3 and 6 (Year 2 values) plus data on outpatient mental health spending from Wells et al. inflated to 1984 and scaled up by 207 million population imply an expenditure of $224 billion on the free care plan in our sites and $178 billion on the 25 percent coinsurance plans. Adjusting for price and usage levels in our sites relative to the nation is necessarily imprecise, but these two values probably bracket the true national figure.

[46]The induced technological change is clearly only a welfare loss if patent protection is at the level to induce the appropriate investment in new products in an unsubsidized market. If there is not enough patent protection, there is no necessary welfare loss from insurance's inducing a too rapid rate of innovation. There appears to be one estimate in the literature of the welfare loss from induced change; Feldstein (1973) attempted to adjust for the willingness of consumers to pay for "higher quality care." There is no empirical way to do this, however, so the magnitude of the true welfare loss is highly problematic. Feldstein's method, although not explicit on the point, in effect ignores true technological change. He implicitly assumes that consumers in earlier years could have purchased "higher quality" medical care, but they chose not to because they faced a higher coinsurance rate and/or had lower incomes. (Alternatively, physician "norms of care" were lower because of the higher coinsurance rate and lower income.) As the renal dialysis example makes clear, however, consumers were simply unable to purchase some medical services in earlier years because they did not exist. In many cases their subsequent existence depended on fundamental scientific advance such as the discovery of DNA and would not have occurred without that advance, despite lower coinsurance or higher incomes. Whether consumers in the 1950's and early 1960's would have purchased such services if they had existed then obviously cannot be answered from actual expenditure data. Feldstein's method also yields an upper bound for the same reason our $60 billion estimate is an upper bound.

centage of premium) are less, and the risk of a large loss is greater. For children, price elasticities for inpatient services are not measurably different from zero, and hence for them there is no measurable moral hazard.

This structure of more extensive insurance for inpatient services has been attacked as misguided, however (Milton Roemer et al., 1975), on the grounds that lack of insurance for outpatient services deters ignorant individuals from seeking care at a time in their illness when they can be treated relatively cheaply. Others have also asserted that the more generous coverage of inpatient services leads physicians to hospitalize patients who could be treated on an outpatient basis, thereby minimizing private but increasing social expenditure.

Analysis of a natural (not randomized) experiment supported the claim that more complete coverage of outpatient expenditure reduced total expenditure (Roemer et al.; L. Jay Helms, Newhouse, and Phelps, 1978), but a prior controlled experimental study testing this hypothesis rejected it (Charles Lewis and Harold Keairnes, 1970; D. B. Hill and James Veney, 1980). At issue is whether outpatient and inpatient services are substitutes or complements.

Our findings decisively reject the hypothesis that increased coverage of outpatient services, holding constant the coverage of inpatient services, will reduce expenditure. As Table 3 shows, the mean expenditure on the individual deductive plan (free inpatient, costly outpatient care) is 20 percent less than the mean on the free care plan (free inpatient, free outpatient care), and the difference is statistically significant ($p < .001$).[47] Disaggregation shows that the outpatient deductible not only reduces outpatient expenditure (Table 2) but, if anything, decreases hospital admissions for adults as well (Table 5). The (possibly) decreased admissions for adults suggests that outpatient and inpatient services are, if anything, complements not substitutes.

In the interests of brevity we summarize four other implications for health insurance coverage (these are discussed at greater length in Manning et al., 1987):

There appears to be little justification for the common practice of group insurance policies' treating emergency room services more generously than physician office visits, because emergency room services are as responsive to plan as physician office visits.[48]

There is no support for the so-called offset hypothesis, namely that more complete coverage of psychotherapy services will reduce total medical costs (or at least not increase them) (W. Follette and Nicholas Cummings, 1967, 1968). The experimental data, however, are not very precise on this question.

The observed lesser coverage of outpatient mental health care relative to all outpatient care would be consistent with a greater plan response for mental health care. Although the estimated plan response is in fact substantially larger for mental health care, the difference with all outpatient care is statistically insignificant.[49]

Well-care services are about as price responsive as other medical services. Although there are other reasons for the common practice of not covering well-care

[47]In the ANOVA results (Table 2), the estimated reduction is 19 percent and the t-statistic is 2.34 ($p < .02$, two-tailed test).

[48]We assume that a presumed lower response to insurance is the reason for greater coverage of emergency room services. The alternative explanations, differential loading charges or asymmetric information, are not particularly plausible as explanations of the better coverage of emergency room services. Asymmetric information (differential knowledge of insurer and insured) is not very relevant to a single insurance plan offered in a group setting unless the service is costly enough to motivate an employment change (which might apply to psychotherapy or certainly costly dental services such as orthodontia). Routine office visits do not match this description. Moreover, asymmetric information may apply to both office and emergency room services. An individual may know that his use of office visits differs from average (whereas the insurer does not) but may also know that his likelihood of an accident differs from average, and the insurer may not.

[49]The estimated ratio of the free to 95 plan expenditures is 233 percent, compared with a 169 percent estimate for medical outpatient care (Manning et al., 1986b).

services as generously as other outpatient services (primarily there is little or no uncertainty and loadings are relatively high), greater price responsiveness is not a reason.

E. *Was It Worth It?*

One question frequently raised about social experimentation is whether its benefits are worth its costs (for example, Orley Ashenfelter, 1986; Robert Haveman, 1986). Because the question concerns the value of information, and because the links from this type of information to actual behavior are generally impossible to establish with any rigor, the question admits of no easy answer (save for the trivial case in which the experiment was so poorly designed or conducted that it produced no information). In other words, any attempt to justify the cost of an experiment is necessarily speculative.

Despite the circumstantial nature of the evidence, we believe that the benefits of this particular experiment greatly exceeded the (current dollar, undiscounted) costs of a little over $80 million ($136 million if put in 1984 dollars, and brought forward to 1984 using a 3 percent real discount rate.[50] Between 1982 and 1984, there was a remarkable increase in initial cost sharing in the United States, at least for hospital services. For example, the number of major companies with first-dollar charges for hospital care rose from 30 to 63 percent in those two years, and the number of such firms with an annual deductible of $200 per person or more rose from 4 percent to 21 percent (Jeff Goldsmith, 1984). Although it is impossible to know how much of this change can be attributed to the experimental results, the initial findings of the experiment were published in December 1981 (Newhouse et al., 1981) and December 1983 (Brook et al., 1983) and given wide publicity in both the general and trade press. In certain instances a direct link between changes in cost sharing and the experimental results can be made.[51]

According to the experimental results, this increase in cost sharing should have decreased demand. Hospital days (excluding deliveries) among the under 65 decreased by 19 million days, or 13 percent, discharges decreased by 8 percent (USDHHS, Series 13, 1984; 1986). We estimate the cost saving from this reduced use to be around $7 billion.[52] Physician office and hospital visits among the under 65 fell 27 million during these two years, but to be conservative we have not taken account of this change in estimating the cost savings.[53]

[50]We have used the GNP deflators to inflate costs.

[51]For example, the Xerox Corporation in 1983 announced an increase in its deductible from $100 per person or $200 per family to 1 percent of earnings per family. It raised coinsurance on hospital and surgical services from 0 to 20 percent. Additionally, it lowered its cap on out-of-pocket expenditures (analogous to the MDE) from 6 percent of earnings to 4 percent of earnings. In a brochure distributed to its employees it said: "According to a study by the Rand Corporation, when consumers are required to increase their share of medical costs, there is a significant decrease in the total amount spent for these services. Furthermore, this study —and other similar studies—does not indicate that the health of the employees was affected adversely by the decrease in costs." Despite the large increase in initial cost sharing, the average coinsurance rate for hospital services nationally only rose from 7.6 to 8.7 percent between 1982 and 1984. This modest change in the average rate may reflect both the lowering of ceilings on out-of-pocket expenditure, as in the Xerox case and the highly skewed distribution of hospital expenditure, which means most expenditure exceeds the initial cost sharing.

[52]The average cost per hospital day in 1984 was $417. This uses the 1983 $368 figure from the American Hospital Association (1984) inflated by 13.3 percent, the change in per day inpatient costs from 1983 to 1984 (American Hospital Association, 1985). Bernard Friedman and Pauly (1981, 1983) have argued that the marginal cost/average cost ratio for hospital services is near one. Hence, a *ceteris paribus* estimate of the savings from decreased use, assuming a marginal cost/average cost ratio of 0.9, is around $7 billion (19 million $\times$ 417 $\times$ 0.9). The American Hospital Association cost per day figure includes the over 65; however, cost per day is not very different for the over 65.

[53]In part, we do not account for such a change because the physician visit rate rose in 1985 to its 1982 value. Thus, the decrease from 1982 to 1984 could have been attributable to chance; alternatively the continued decrease in hospital care in 1985 (another 7.1 percent decrease in patient-days, USDHHS, 1987) may have led to a substitution of outpatient use. Data on physician visits are from the *National Health Interview Survey* (USDHHS, Series 10, 1985; 1986).

If all the changes in patient-days were attributable to the increased cost sharing, and if all the increase in cost sharing is due to the publication of experimental results, and if the benefits of the foregone use were negligible, as our results suggest, the experiment paid for itself in about a week (.136/7)(52)![54] It is clear that these assumptions overstate the benefits of the experiment, yet it is equally clear that the assumptions can be greatly relaxed and still yield the result that the experiment was worth it. Moreover, we have ignored any benefits to countries other than the United States, and any benefits from the decrease in physician visits or changes in dental or mental health coverage or emergency room coverage. We have also ignored any benefits from the results of the HMO portion of the experiment, although HMO's market share has been expanding rapidly from a period just before and subsequent to our first article describing the HMO results (Manning et al., 1984a). Finally, we have ignored the value of the public use files to future research efforts.[55]

Implicit in our conclusions is the assumption that one could not reduce uncertainty with nonexperimental data to the satisfaction of those making decisions about cost sharing. We believe this is likely to be true, because of the wide range of nonexperimental estimates of insurance elasticity cited in the introduction, the difficulty of inferring health status effects from nonexperimental data, and the temporal proximity of the changes in cost sharing to the publication of the experimental results (many of the nonexperimental results had been in the literature for a decade, during which time cost sharing had, if anything, decreased). Thus, we think it highly plausible that the benefits of this endeavor were indeed worth its costs.

[54]The negligible benefits assumption relies on the observation that cost sharing for hospital services was near zero in 1982 and that there were no measurable health benefits outside the dental area for the middle-class employees who would have been the dominant group for whom the cost sharing changed.

[55]The public use files can be ordered from Publications Department, The Rand Corporation, 1700 Main Street, Santa Monica, CA 90406-2138.

F. *On Experimentation in Economics*

Econometric and economics texts often have a statement near the beginning that experimentation is not nearly as possible in economics as it is in the physical sciences. Perhaps the degree of difference is not as great as many think. Well-designed and executed field and laboratory experiments are feasible and can add substantially to the body of knowledge (Walter Heller, 1975; Charles Plott, 1982).[56] We hope this example will encourage others to ask whether an experiment is practical or feasible when approaching empirical questions.

[56]For other views of field experiments see Jerry Hausman and David Wise (1985) and Robert Ferber and Werner Hirsch (1978).

REFERENCES

Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, December 1963, *53*, 941–73.

______, *Essays in the Theory of Risk-Bearing*, Chicago: Markham, 1971.

______, *Optimal Insurance and Generalized Deductibles*, Publ. No. R-1108-OEO, Santa Monica: Rand Corporation, 1973.

______, "Two Notes on Inferring Long Run Behavior from Social Experiments," Publ. No. P-5546, Rand Corporation, 1975.

Ashenfelter, Orley, "Book Review of *The National Supported Work Demonstration*," *Journal of Economic Literature*, September 1986, *24*, 1268–1270.

Bailit, Howard et al., "Does More Generous Dental Insurance Coverage Improve Oral Health?," *Journal of the American Dental Association*, May 1985, *110*, 701–07.

Barer, Morris L., Evans, Robert G. and Stoddart, Gregory L., "Controlling Health Care Costs by Direct Charges to Patients: Snare or Delusion?," Occasional Paper No. 10, Toronto: Ontario Economic Council, 1979.

Baumol, William J. and Bradford, David F., "Optimal Departures from Marginal Cost Pricing," *American Economic Review*, June 1970, *60*, 263–83.

Beck, R. G., "The Effects of Co-Payment on the Poor," *Journal of Human Resources*, Winter 1974, *9*, 129–42.

Brook, Robert H. et al., "Overview of Adult Health Status Measures Fielded in Rand's Health Insurance Study," *Medical Care*, 1979, Suppl., *17*, 1–131.

______ **et al.**, "Does Free Care Improve Adults' Health? Results from a Randomized Controlled Trial," *New England Journal of Medicine*, December 8, 1983, *309*, 1426–34.

______ **et al.**, *The Effect of Coinsurance on the Health of Adults*, Publ. No. R-3055-HHS, Santa Monica: Rand Corporation, 1984.

Clasquin, Lorraine A., *Mental Health, Dental Services, and Other Coverage in the Health Insurance Study*, Publ. No. R-1216-OEO, Santa Monica: Rand Corporation, 1973.

______ **and Brown, Marie E.**, *Rules of Operation for the Rand Health Insurance Study*, Publ. No. R-1602-HEW, Santa Monica: Rand Corporation, 1977.

Colle, Ann D. and Grossman, Michael, "Determinants of Pediatric Care Utilization," *Journal of Human Resources*, 1978, Suppl., *13*, 115–58.

Culyer, Anthony J., "The Nature of the Commodity 'Health Care' and Its Efficient Allocation," *Oxford Economic Papers*, July 1971, *23*, 189–211.

______, *Need and the National Health Service* Totowa: Rowman and Littlefield, 1976.

______, "Measuring Health: Lessons for Ontario," Ontario Economic Council Research Studies No. 14, University of Toronto, 1978.

Davies, Allyson et al. "Consumer Acceptance of Prepaid and Fee-for-Service Medical Care: Results from a Randomized Trial," *Health Services Research*, August 1986, *21*, 429–52.

Davis, Karen and Russell, Louise B., "The Substitution of Hospital Outpatient Care for Inpatient Care," *Review of Economics and Statistics*, May 1972, *54*, 109–20.

Duan, Naihua, "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association*, September 1983, *78*, 605–10.

______ **et al.**, "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Economic and Business Statistics*, April 1983, *1*, 115–26 [also R-2754-HHS, Rand Corporation, 1982].

______ **et al.**, "Choosing Between the Sample-Selection Model and the Multi-Part Model," *Journal of Business and Economic Statistics*, July 1984, *2*, 283–89.

Ellis, Randall P., "Rational Behavior in the Presence of Coverage Ceilings and Deductibles," *Rand Journal of Economics*, Summer 1986, *17*, 158–75.

Evans, Robert G., *Strained Mercy*: *The Economics of Canadian Health Care*, Toronto: Butterworths, 1984.

Feldstein, Martin S., "Hospital Cost Inflation: A Study of Nonprofit Price Dynamics," *American Economic Review*, December 1971, *61*, 853–72.

______, "The Welfare Loss of Excess Health Insurance," *Journal of Political Economy*, March/April 1973, *81*, 251–80.

______, "Quality Change and the Demand for Hospital Care," *Econometrica*, October 1977, *45*, 1681–702.

______ **and Allison, Elizabeth**, "Tax Subsidies of Private Health Insurance," in *The Economics of Federal Subsidy Programs*, Pt. 8, papers submitted to the Subcommittee on Priorities and Economy in Government of the Joint Economic Committee, 93rd Congress, 2nd session, July 1974.

______ **and Friedman, Bernard**, "Tax Subsidies, the Rational Demand for Insurance, and the Health Care Crisis," *Journal of Public Economics*, April 1977, *7*, 155–78.

Ferber, Robert and Hirsch, Werner, "Social Experimentation and Economic Policy: A Survey," *Journal of Economic Literature*, December 1978, *16*, 1379–404.

Follette, W. and Cummings, Nicholas A., "Psychiatric Services and Medical Utilization in a Prepaid Health Plan Setting," *Medical Care*, January 1967, *5*, 26–35.

______ **and** ______, "Psychiatric Services and Medical Utilization in a Prepaid Health Plan Setting, Part 2," *Medical Care*, January 1968, *6*, 31–41.

Friedman, Bernard and Pauly, Mark, "Cost Functions for a Service Firm with Variable Quality and Stochastic Demand," *Review of Economics and Statistics*, November 1981, *63*, 610–24.

______ **and** ______, "A New Approach to Hospital Cost Functions and Some Issues

in Revenue Regulation," *Health Care Financing Review*, March 1983, *4*, 105–14.

Goddeeris, John and Weisbrod, Burton, "What We Don't Know about Why Health Expenditures Have Soared: Interaction of Insurance and Technology," *Mount Sinai Journal of Medicine*, November 1985, *52*, 685–91.

Goldman, Fred and Grossman, Michael, "The Demand for Pediatric Care: An Hedonic Approach," *Journal of Political Economy*, April 1978, *86*, 259–80.

Goldsmith, Jeff, "Death of a Paradigm: The Challenge of Competition," *Health Affairs*, Fall 1984, *3*, 5–19.

Hausman, Jerry and Wise, David, *Social Experimentation*, Chicago: University of Chicago Press, 1985.

Haveman, Robert, "Social Experimentation and *Social Experimentation*," *Journal of Human Resources*, Fall 1986, *21*, 586–605.

Helms, L. Jay, Newhouse, Joseph, P. and Phelps, Charles E., "Copayments and the Demand for Medical Care: The California Medicaid Experience," *Bell Journal of Economics*, Spring 1978, *9*, 192–208.

Heller, Walter, W., "What's Right with Economics," *American Economic Review*, March 1975, *65*, 1–26.

Hill, D. B. and Veney, James E., "Kansas Blue Cross/Blue Shield Outpatient Benefits Experiment," *Medical Care*, March-April 1980, *8*, 143–58.

Huber, P. J., "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 1967, 221–33.

Keeler, Emmett B. and Rolph, John E., *The Demand for Episodes of Medical Services: Interim Results from the Health Insurance Experiment*, R-2829-HHS, Santa Monica: Rand Corporation, December 1982.

______, Morrow, Daniel and Newhouse, Joseph P., "The Demand for Supplementary Health Insurance, or Do Deductibles Matter?," *Journal of Political Economy*, August 1977, *85*, 789–802.

______, Newhouse, Joseph, P. and Phelps, Charles E., "Deductibles and the Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertainty," *Econometrica*, April 1977, *45*, 641–56.

______ et al., "How Free Care Reduced Hypertension of Participants in the Rand Health Insurance Experiment," *Journal of the American Medical Association*, October 11, 1985, *154*, 1926–31.

______ et al., *The Demand for Episodes of Medical Treatment*, Publ. No. R-3454-HHS, Santa Monica: Rand Corporation, 1987.

Leibowitz, Arleen et al., "Effect of Cost Sharing on the Use of Medical Services by Children: Interim Results from a Randomized Controlled Trial," *Pediatrics*, May 1985, *75*, 942–51.

Levit, Katherine R. et al., "National Health Expenditures, 1984," *Health Care Financing Review*, Fall 1985, *7*, 1–35.

Lewis, Charles E. and Keairnes, Harold, "Controlling Costs of Medical Care by Expending Insurance Coverage," *New England Journal of Medicine*, June 18, 1970, *292*, 1405–12.

Lindsay, Cotton M., "Medical Care and the Economics of Sharing," *Economica*, November 1969, *36*, 351–62.

Lohr, Kathleen N. et al., "Use of Medical Care in the Rand Care in the Rand Health Insurance Experiment: Diagnosis- and Service-Specific Analyses," *Medical Care*, September 1986, Suppl.

Long, Stephen H., Settle, Russell F. and Stuart, Bruce C., "Reimbursement and Access to Physicians' Services under Medicaid," *Journal of Health Economics*, September 1986, *5*, 235–52.

Luft, Harold S., *Health Maintenance Organizations*, New York: Wiley & Sons, 1981.

McCombs, Jeffrey S., "Physician Treatment Decisions in a Multiple Treatment Model: The Effects of Physician Supply," *Journal of Health Economics*, August 1984, *3*, 155–71.

Maddala, G. S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Manning, Willard G., "Estimating Demand Functions with Data from Declining Block Tariffs," draft, June 1986.

______, Duan, Naihua and Rogers, William H., *Monte Carlo Evidence on the Choice between Sample Selection and Two-Part Models*, Santa Monica: Rand Corpora-

tion, forthcoming.

______ **et al.**, (1984a) "A Controlled Trial of the Effect of a Prepaid Group Practice on Use of Services," *New England Journal of Medicine*, June 7, 1984, *310*, 1505–10 [also R-3029-HHS, Rand Corporation, December 1985].

______ **et al.**, (1984b) "Cost Sharing and the Demand for Ambulatory Mental Health Services," *American Psychologist*, October 1984, *39*, 1090–100.

______ **et al.**, "The Demand for Dental Care: Evidence from a Randomized Trial in Health Insurance," *Journal of the American Dental Association*, June 1985, *110*, 895–902.

______ **et al.**, (1986a) *Health Insurance and the Demand for Dental Care: Evidence from a Randomized Experiment*, Publ. No. R-3225-HHS, Santa Monica: Rand Corporation, 1986.

______ **et al.**, (1986b) "How Cost Sharing Affects the Use of Ambulatory Mental Health Services," *Journal of the American Medical Association*, 1986, *256*, 1930–34.

______ **et al.**, *Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment*, Publ. No. R-3476-HHS, Santa Monica: Rand Corporation, 1987.

Mehran, F., "Variance of the MVUE for the Log Normal Mean," *Journal of the American Statistical Association*, 1973, *68*, 726–27.

Metcalf, Charles E., "Making Inferences from Controlled Income Maintenance Experiments," *American Economic Review*, June 1973, *63*, 478–483.

Morris, Carl N., "A Finite Selection Model for Experimental Design of the Health Insurance Study," *Journal of Econometrics*, September 1979, *11*, 43–61.

Musgrave, Richard, *The Theory of Public Finance*, New York: McGraw-Hill, 1959.

Newhouse, Joseph P., "A Design for a Health Insurance Experiment," *Inquiry*, March 1974, *11*, 5–27.

______, "Medical Care Expenditure: A Cross-National Survey," *Journal of Human Resources*, Winter 1977, *12*, 115–25.

______, "Insurance Benefits, Out-of-Pocket Payments, and the Demand for Medical Care: A Review of the Literature," *Health and Medical Care Services Review*, July-August 1978, *1*, 3–15.

______, "The Demand for Medical Care Services: A Retrospect and Prospect," in J. van der Gaag and M. Perlman, eds., *Health, Economics, and Health Economics*, Amsterdam: North-Holland, 1981.

______, "Are Medical Care Costs Too High?," *Journal of Dental Education*, November 1984, *48*, 587–90.

______ **and Phelps, Charles E.**, "Price and Income Elasticities for Medical Care Services," in M. Perlman, ed., *The Economics of Health and Medical Care*, New York: Wiley & Sons, 1974.

______ **and** ______, "New Estimates of Price and Income Elasticities of Medical Care Services," in R. N. Rosett, ed., *The Role of Health Insurance in the Health Services Sector*, Universities-National Bureau Conference Series No. 27, New York, 1976.

______, ______, **and Marquis, M. Susan**, "On Having Your Cake and Eating It Too: Econometric Problems in Estimating the Demand for Health Services," *Journal of Econometrics*, August 1980, *13*, 365–90.

______ **et al.**, "Design Improvements in the Second Generation of Social Experiments: The Health Insurance Study," *Journal of Econometrics*, September 1979, *11*, 117–29.

______ **et al.**, "Some Interim Results from a Controlled Trial of Cost Sharing in Health Insurance," *New England Journal of Medicine*, December 17, 1981, *305*, 1501–07 [see R-2847-HHS, Rand Corporation, 1982].

______ **et al.**, "The Findings of the Rand Health Insurance Experiment: A Response to Welch et al.," *Medical Care*, February 1987, *25*, 157–79.

O'Grady, Kevin F. et al., "The Impact of Cost Sharing on Emergency Department Use," *New England Journal of Medicine*, August 22, 1985, *313*, 484–90.

Pauly, Mark V., *Medical Care at Public Expense*, New York: Prager, 1971.

______, "Taxation, Health Insurance, and Market Failure," *Journal of Economic Literature*, June 1986, *24*, 629–75.

Phelps, Charles E. and Newhouse, Joseph P., "Effects of Coinsurance: A Multivariate Analysis," *Social Security Bulletin*, June 1972, *35*, 20–29.

______ and ______, "Coinsurance, the Price of Time, and the Demand for Medical Services," *Review of Economics and Statistics*, August 1974, *56*, 334–42.

Plott, Charles R., "Industrial Organization Theory and Experimental Economics," *Journal of Economic Literature*, December 1982, *20*, 1435–527.

Ramsey, Frank, "A Contribution to the Theory of Taxation," *Economic Journal*, March 1927, *37*, 47–61.

Roemer, Milton J. et al., "Copayments for Ambulatory Care: Penny-wise and Pound-foolish," *Medical Care*, 1975, *13*, 475–66.

Rogers, William H., and Newhouse, Joseph P., "Measuring Unfiled Claims in the Health Insurance Experiment," in L. Burstein et al., eds., *Collecting Evaluation Data: Problems and Solutions*, Beverly Hills: Sage, 1985, 121–33.

Rosett, Richard N. and Huang, Lien Fu, "The Effect of Health Insurance on the Demand for Medical Care," *Journal of Political Economy*, March/April 1973, *81*, 281–305.

Rothschild, Michael and Stiglitz, Joseph, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quarterly Journal of Economics*, November 1976, *90*, 629–50.

Scitovsky, Anne A. and McCall, Nelda M., "Coinsurance and the Demand for Physician Services: Four Years Later," *Social Security Bulletin*, May 1977, *40*, 19–27.

______ **and Snyder, Nelda M.**, "Effect of Coinsurance on Use of Physician Services," *Social Security Bulletin*, June 1972, *35*, 3–19.

Searle, S. R., *Linear Models*, New York: Wiley & Sons, 1971.

Sloss, Elizabeth M. et al., "Effect of a Health Maintenance Organization on Physiologic Health: Results from a Randomized Trial," *Annals of Internal Medicine*, January 1987, *106*, 130–38.

Tobin, James, "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, January 1958, *26*, 24–36.

Valdez, R. Burciaga, *The Effects of Cost Sharing on the Health of Children*, Publ. No. R-3270, Santa Monica: Rand Corporation, 1986.

______ **et al.**, "The Consequences of Cost Sharing for Children's Health," *Pediatrics*, May 1985, *75*, 957–61.

Van de Ven, Wynand P. and van Praag, Bernard M., (1981a) "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection," *Journal of Economics*, November 1981, *17*, 229–52.

______ **and** ______, (1981b) "Risk Aversion of Deductibles in Private Health Insurance," in J. van der Gaag and M. Perlman, eds., *Health, Economics, and Health Economics*, Amsterdam: North-Holland, 1981.

Waldo, Daniel and Lazenby, Helen C., "Demographic Characteristics and Health Care Use and Expenditures by the Aged in the United States, 1977–1984," *Health Care Financing Review*, Fall 1984, *6*, 1–29.

______, **Levit, Katherine R. and Lazenby, Helen C.**, "National Health Expenditures, 1985," *Health Care Financing Review*, Fall 1986, *8*, 1–21.

Ware, John E. et al., *Conceptualization and Measurement of Health*, Vol. I, R-1987/1-HEW, Santa Monica: Rand Corporation, May 1980.

______ **et al.**, "Comparison of Health Outcomes at a Health Maintenance Organization with Those of Fee-for-Service Care," *The Lancet*, May 3, 1986, *848*, 1017–22.

Wells, Kenneth B. et al., *Cost Sharing and the Demand for Ambulatory Mental Health Services*, Publ. No. R-2960-HHS, Santa Monica: Rand Corporation, 1982.

Zeckhauser, Richard J., "Medical Insurance: A Case Study of the Trade-Off between Risk Spreading and Appropriate Incentives," *Journal of Economic Theory*, March 1970, *2*, 10–26.

American Hospital Association, "1984 Hospital Cost and Utilization Trends," *Economic Trends*, Spring 1985, 1.

______, *Hospital Statistics*, 1984 Edition, Chicago: AH Association, 1984.

U.S. Department of Health and Human Services, "Current Estimates from the *National Interview Health Survey*," Series 10: 150, 1985; 156, 1986, Washington: USDHHS.

______, "Utilization of Short Stay Hospitals," Series 13: 78, 1984; 84, 1986; 90, 1987; Washington: USDHHS.

Job Duration, Seniority, and Earnings

By Katharine G. Abraham and Henry S. Farber*

An important stylized fact about labor markets is that workers with longer seniority with their current employer have higher earnings than other workers with the same total labor market experience. This study shows that the measured positive cross-sectional return to seniority is largely a statistical artifact due to the correlation of seniority with an omitted variable representing the quality of the worker, job, or worker-employer match. The implication is that earnings do not, in fact, rise very much with seniority.

It is a commonly accepted empirical finding that workers with more seniority on their current job earn more than other workers with the same total labor market experience. The standard explanations for a positive correlation between seniority and earnings are based on the existence of implicit employment contracts under which earnings grow with time on the job in order to provide workers with appropriate incentives regarding turnover and/or effort. For example, if a job involves investment in firm-specific training, then it may be optimal for workers and employers to structure implicit employment agreements such that compensation is deferred until late in the job so that workers will not quit (taking their specific capital with them).[1] Another possible motivation for such a deferral arrangement exists where effort is important. The promise of eventual compensation in excess of the opportunity value of time provides the worker with an incentive to exert the appropriate level of effort on the job. A worker who left or whose performance fell below agreed-upon standards and in consequence was fired would lose the opportunity to enjoy the benefits of the high deferred wages.[2]

The firm-specific human capital explanation and effort-incentive wage deferral explanations differ in their implications regarding the relationship between earnings and productivity growth over the work life.[3] However, they share the implication not only that the seniority-earnings profile will be upward sloping, but also that there will be a positive return to seniority even after controlling for total labor market experience.

The empirical support for these views of the labor market rests entirely on the positive cross-sectional association between seniority and earnings, but this is not sufficient evidence to establish that earnings rise with seniority. An alternative interpretation of the

*The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, DC 20036, and Department of Economics, E52-252f, MIT, Cambridge MA 02139, respectively. We thank Joe Altonji for providing us with the data, Jim Powell for providing valuable econometric advice, and Leslie Sundt for able research assistance in preparing the data for analysis. Useful comments on an earlier version of this paper were received from Altonji, Jerry Hausman, Steve Nickell, Bob Topel, and participants in numerous workshops. Farber received support from the National Science Foundation. The views expressed here are our own and should not be attributed to the trustees, officers, or staff members of the institutions with which we are affiliated.

[1]See Gary Becker (1964), Jacob Mincer (1974), and Dale Mortensen (1978) for discussions of investment in firm-specific training. Mincer and Boyan Jovanovic (1981) present an analysis of the relationships among seniority, mobility, and earnings that relies on investment in specific human capital.

[2]See W. Kip Viscusi (1980), Becker and George Stigler (1974), and Edward Lazear (1979) for models in which wage deferral provides this sort of incentive for workers. George Akerlof and Lawrence Katz (1986) argue that such deferral arrangements cannot, in fact, yield efficient effort levels unless the present value of promised earnings at the start of the job exceeds the present value of earnings on the next-best alternative job.

[3]James Medoff and Abraham (1980, 1981) and Abraham and Medoff (1982) offer evidence on the relationship between seniority-related earnings growth and seniority-related productivity growth.

cross-sectional evidence is based on the idea that workers who are 1) better workers, 2) in better jobs, or 3) in better worker-employer matches earn more throughout their jobs and also stay on their jobs longer. It is straightforward to show that the distribution of seniority in a cross section has a higher mean for workers on longer jobs. Thus, as long as those workers who earn more from the start have longer average completed job durations, the ordinary least squares estimate of the return to seniority in a cross-section earnings function is biased upward.

The key testable link in this argument is that workers in long jobs earn more from the start than observationally equivalent workers in short jobs. Why might this be true? First, some workers may be both more stable and more productive than others. To the extent that there are turnover and training costs, stability per se can raise an employee's value to the firm. Second, some employers may choose to pay higher wages than others (some jobs are "better" than others), so that workers are unlikely either to shirk or to quit.[4] Finally, some worker/job matches may be better than others, in the sense that the worker is more productive in those matches than in other possible matches, so that the specific value of the match is shared between the employer and the worker, and the match is less likely to be broken off.[5] To the extent that worker, job, and/or match quality are unmeasured, they represent omitted variables in a cross-section earnings regression.

Here we develop a simple stochastic model of earnings determination that illustrates the upward bias in the estimated return to seniority just described and provides the basis for two related approaches to correcting this bias. These approaches are implemented using data from the *Panel Study of Income Dynamics*. The results of this empirical analysis reveal 1) that workers on longer jobs earn more in every year on the job, and 2) that much of the apparent cross-section return to seniority reflects omitted variable bias.

I. Completed Job Duration in the Earnings Function

Suppose that the earnings of a particular worker i on job j in year t can be written

$$(1)\quad \ln W_{ijt} = \beta_1 S_{ijt} + \beta_2 EXP_{ij} + \mu_{ij} + \eta_{ijt},$$

where W = hourly earnings, S = current seniority (tenure), β_1 = return to seniority, EXP = pre-job experience, β_2 = return to pre-job experience, μ = a person/job-specific error term representing the excess of earnings enjoyed by this person on this job over and above the earnings that could be expected by a randomly selected person/job combination, and η = a person/job/time-period-specific error term.

For simplicity of exposition, other factors that might influence earnings are omitted from the theoretical discussion and all variables are assumed to be measured as deviations from their means. In this formulation, μ_{ij} captures the net influence of three unobservables on hourly earnings: unobserved person quality, unobserved job quality, and unobserved match quality. The error μ_{ij} is assumed to be fixed over the course of a job and may be correlated with S and EXP. The error η_{ijt} is assumed to be orthogonal to S, EXP, and μ.

In equation (1), β_2 represents the returns to experience per se, including the returns to general human capital and any other growth in earnings that occurs automatically with time in the labor market. Additionally, earnings are likely to grow with experience because more experienced workers typically end

[4] The efficiency wage literature suggests various possible reasons why some employers might pay higher wages than others to workers of equal quality. These include differences in the costs of turnover, differences in the costs of monitoring worker shirking, and differences in the value of worker loyalty. Carl Shapiro and Joseph Stiglitz (1984) and Jeremy Bulow and Lawrence Summers (1986) present formal efficiency wage models. Janet Yellen (1984), Stiglitz (1984), and Katz (1986) survey the literature.

[5] Mortensen and Jovanovic (1979) both present theoretical analyses of the effects of heterogeneous match quality that have these implications.

up in better jobs and/or better matches.[6] This will be reflected in higher values of μ_{ij} as EXP_{ij} increases. Let the relationship between pre-job experience and μ be approximated by

$$\mu_{ij} = \alpha EXP_{ij} + \phi_{ij}, \tag{2}$$

where α captures the growth in μ_{ij} with experience and ϕ_{ij} is the component of μ_{ij} that is uncorrelated with EXP_{ij}. Substituting into equation (1) yields

$$\ln W_{ijt} = \beta_1 S_{ijt} + (\beta_2 + \alpha) EXP_{ij} + \phi_{ij} + \eta_{ijt}, \tag{3}$$

where β_1 is the total return to seniority and $\beta_2 + \alpha$ is the total return to pre-job experience including both the return to experience per se and systematic returns to search. The net return to seniority is appropriately defined as the excess of the growth in earnings on a given job over and above the total returns to general labor market experience, or $\beta_1 - (\beta_2 + \alpha)$.

In practice, earnings functions are generally estimated using cross-section data and ϕ_{ij} is not observable. The standard cross-section earnings equation is

$$\ln W_{ijt} = b_1 S_{ijt} + b_2 EXP_{ij} + v_{ijt}, \tag{4}$$

where v is the estimating equation error. Omission of ϕ_{ij} from equation (4) means that b_1 and b_2 may be biased estimators of the return to seniority, β_1, and the return to experience, $\beta_2 + \alpha$. Deriving the expected values of b_1 and b_2 requires knowing more about the relationship between ϕ_{ij} and S_{ijt}. In particular, we need to know the partial relationship between ϕ_{ij} and S_{ijt}, holding EXP_{ij} constant.[7]

It was argued above that good workers and workers in good jobs or good matches are likely to stay on their jobs longer. Formally, this implies that the completed duration of jobs is positively related to μ_{ij}. Let this relationship be expressed as

$$D_{ij} = \gamma \mu_{ij} + \varepsilon_{ij}, \tag{5}$$

where D_{ij} is the completed length of the current job, γ is a parameter that summarizes the relationship between D and μ, and ε_{ij} captures the variation in completed job duration that cannot be linked to the earnings advantage associated with worker, job, and/or match quality.[8] Substituting from equation (2),

$$D_{ij} = \alpha\gamma EXP_{ij} + \gamma\phi_{ij} + \varepsilon_{ij}. \tag{6}$$

Holding initial experience constant, completed job duration is positively related to ϕ_{ij}.

What does this tell us about the relationship between ϕ_{ij} and S_{ijt}? In a cross section of individuals, those with longer current seniority are likely to be on longer-lasting jobs. More formally, if each year of any given job is equally likely to be represented in the cross-section sample of observations used to estimate the earnings function, then on average the observed seniority on the job will be halfway through the job so that

$$\begin{aligned} S_{ijt} &= 1/2 \times D_{ij} + \xi_{ijt} \\ &= 1/2 \times \alpha\gamma EXP_{ij} + 1/2 \times \gamma\phi_{ij} \\ &\quad + 1/2 \times \xi_{ij} + \xi_{ijt}, \end{aligned} \tag{7}$$

where ξ_{ijt} is a random variable with zero mean. Thus, the existence of a positive relationship between ϕ and D, holding pre-job experience constant, implies the existence of a positive relationship between ϕ and S,

[6]See Kenneth Burdett (1978), Jovanovic, and Robert Topel (1986).

[7]By construction, ϕ_{ij} is uncorrelated with EXP_{ij}. However, if S_{ijt} and EXP_{ij} are correlated, omitting ϕ_{ij} may bias both b_1 and b_2.

[8]There is a large literature that estimates turnover probabilities as a function of wage rates. Donald Parsons (1977) and Richard Freeman and Medoff (1984) present surveys of some of this work. Job duration can be viewed as an inverse turnover measure. The prediction that job duration should be positively related to μ is consistent with evidence that, *ceteris paribus*, turnover rates are negatively related to wages.

holding pre-job experience constant. The distribution of ξ_{ijt} will vary depending upon the completed length of the job. However, its mean is always zero, as are its covariances with μ_{ij}, ϕ_{ij}, D_{ij}, and EXP_{ij}.

Using the relationships in equations (4) to (7), it can be shown that the expected value of the seniority coefficient, b_1, in equation (3) is

$$(8)\quad E(b_1)=\beta_1+\left[\gamma\times\mathrm{var}(EXP_{ij})\times\mathrm{var}(\phi_{ij})\right] \Big/\Big[2\times\big[\mathrm{var}(S_{ijt})\times\mathrm{var}(EXP_{ij}) - \mathrm{cov}^2(S_{ijt}, EXP_{ij})\big]\Big].$$

$E(b_1)$ is larger than β_1 provided γ (the coefficient summarizing the relationship between D and μ) is positive. The expected value of the pre-job experience coefficient, b_2, in equation (4) is

$$(9)\qquad E(b_2)=\beta_2+\alpha K_1,$$

where the multiplicand of α is

$$(10)\quad K_1=\Big[\mathrm{var}(EXP_{ij})\times\big\{\mathrm{var}(S_{ijt}) - 1/4\times\mathrm{var}(D_{ij})+1/4\times\mathrm{var}(\varepsilon_{ij})\big\}\Big] \Big/\Big[\mathrm{var}(S_{ijt})\times\mathrm{var}(EXP_{ij}) - \mathrm{cov}^2(S_{ijt}, EXP_{ij})\Big].$$

It is straightforward to show that K_1 is positive but less than one so that the experience coefficient (b_2) is a downward-biased estimate of $\beta_2+\alpha$. Therefore, b_1-b_2 has an expected value greater than the true net return to seniority, $\beta_1-(\beta_2+\alpha)$, both because b_1 is an upward-biased estimate of β_1 and, secondarily, because b_2 is a downward-biased estimate of $\beta_2+\alpha$.

One approach to correcting the estimates of b_1 and b_2 from equation (4) is to find an instrument for the seniority variable.[9] Equation (7) suggests a suitable instrument: ξ_{ijt}, which equals $S_{ijt}-1/2\times D_{ij}$. By construction, this instrument is correlated with seniority but orthogonal to everything else in the equation. The vector of parameters estimated using ξ as an instrument for S equals

$$(11)\qquad \bar{b}=(Z'X)^{-1}Z'W,$$

where $\bar{b}$ is a 2×1 vector containing $\bar{b}_1$ and $\bar{b}_2$, X is the $n\times 2$ matrix containing values of S_{ijt} and EXP_{ij}, Z is the $n\times 2$ matrix containing values of ξ_{ijt} and EXP_{ij}, and W is the $n\times 1$ matrix of observations on the log wage. The instrumental variables estimator of the seniority coefficient, $\bar{b}_1$, has expected value β_1. In addition, the instrumental variables estimator of the experience coefficient, $\bar{b}_2$, has expected value $\beta_2+\alpha$. Finally, $\bar{b}_1-\bar{b}_2$ is an unbiased estimate of the net return to seniority.

The key assumption underlying the proposed instrumental variables estimator is that observed seniority in a cross section is, on average, half completed job duration. This assumption may not hold exactly. For example, a sample might include a disproportionate number of workers near the start of their work lives who are, on average, closer to the start of their jobs than they are to the end. In this case, the deviation of seniority from one-half completed duration may be correlated with duration and thus not be a valid instrument for seniority. However, if some more general relationship between seniority and completed job duration holds, such as

$$(12)\qquad S_{ijt}=\omega_0+\omega_1 D_{ij}+\xi_{ijt},$$

then, regardless of the values of ω_0 and ω_1, the residual from this regression is an appropriate instrument for seniority. Using estimates of these residuals to calculate the instrument yields consistent estimates of the returns to experience and seniority.

An alternative approach to removing the upward bias in the estimated total return to seniority is to control explicitly for completed job duration in the earnings equation.

[9]Joseph Altonji and Robert Shakotko's (1987) analysis of the return to seniority uses this approach, though with a somewhat different instrument than ours.

Intuitively, the tenure coefficient is biased only because S_{ijt} is associated with D_{ij}, which in turn is correlated with ϕ_{ij}. This suggests that controlling for D_{ij} should eliminate the upward bias in the estimated tenure coefficient. Augmenting the standard cross-section earnings equation by adding D_{ij} as an explanatory variable yields

$$(13)\quad \ln W_{ijt} = \hat{b}_1 S_{ijt} + \hat{b}_2 EXP_{ij} + \hat{b}_3 D_{ij} + \pi_{ijt},$$

where π is the estimating equation error.

It can readily be shown that $E(\hat{b}_1)$ equals β_1 so that $\hat{b}_1$ is an unbiased estimator of the gross return to seniority. The experience coefficient in equation (13) has expected value:

$$(14)\quad E(\hat{b}_2) = \beta_2 + \alpha K_2,$$

where

$$(15)\quad K_2 = \left[\mathrm{var}(EXP_{ij}) \times \mathrm{var}(\varepsilon_{ij})\right] / \left[\mathrm{var}(D_{ij}) \times \mathrm{var}(EXP_{ij}) - \mathrm{cov}^2(D_{ij}, EXP_{ij})\right].$$

The value of K_2 is positive but less than one, which means that $\hat{b}_2$ is an underestimate of $\beta_2 + \alpha$. Thus, so long as α is positive, $\hat{b}_1 - \hat{b}_2$ is an upward-biased estimate of the net return to seniority $(\beta_1 - [\beta_2 + \alpha])$. None of these results are sensitive to the precise form of the linear relationship between seniority and completed duration, and this approach can be used to bound the true return to seniority.

A major attraction of the augmented OLS approach is that it provides a direct estimate of the relationship between completed job duration and earnings, $\hat{b}_3$. This is an indicator of the importance of the relationship of individual, job, and/or match heterogeneity with earnings through job duration. In addition, it serves as a useful device for investigating the underlying hypothesis that better workers, jobs, or matches are associated with higher earnings *throughout* the job. In terms of the earlier analysis, it can be shown that

$$(16)\quad E(b_3) = (1 - K_2)/\gamma,$$

where K_2 is as just defined. Thus, $\hat{b}_3$ is a downward-biased estimate of $1/\gamma$, the coefficient of the regression of μ_{ij} on D_{ij}.

The preceding discussion assumes that individual-, job-, and match-specific earnings components all have the same incremental association with job duration. A more general specification would allow for separate earnings increments associated with unobserved individual characteristics and job/match characteristics. Given that any bias in the estimated return to seniority attributable to omitted earnings components is mediated through the relationship of those omitted earnings components with completed job duration, all of the results just described for the simple model are unaffected by a generalization to multiple unobserved earnings components.

Using the identity that links total experience with pre-job experience and seniority ($EXP_{ijt} = EXP_{ij} + S_{ijt}$) and ignoring second-order terms, valid inferences regarding the net return to seniority can be drawn from an earnings function specification that includes either pre-job experience or total experience along with seniority. Where pre-job experience is used, the net return to seniority $(\beta_1 - [\beta_2 + \alpha])$ is calculated as the difference between the coefficient on seniority and the coefficient on pre-job experience. Where total experience is used, the net return to seniority is simply the coefficient on seniority. To facilitate discussion regarding the net return to seniority, the empirical analysis proceeds using total experience.[10]

[10]For the instrumental variables estimator, this requires the minor modification that total experience (and its square, if included) must be instrumented for along with seniority. The natural additional instruments are pre-job experience (and its square). In a model without squared terms, the IV estimator of the specification using pre-job experience and the IV estimator using total experience with pre-job experience as an ad-

II. Estimating Completed Job Duration

The first step in implementing the analysis described in the preceding section is to derive a measure of completed job duration. The *Panel Study of Income Dynamics* (*PSID*) is used in the empirical analysis. Unfortunately, there is a general problem with virtually all longitudinal data sets, including the *PSID*, when information is required on the completed duration of a spell of any kind. This is that the individuals are followed for only a limited period of time so that there are likely to be many jobs which do not end by the date at which the individual is last observed. Some procedure must be used to impute completed durations to these jobs.

We take the approach of estimating a parametric model of job duration that accounts for the censoring of duration in those jobs for which the end is not observed. This model is then used to compute an estimate of expected completed job duration *conditional* on the job lasting at least as long as the last observed seniority level. This estimate is used as the measure of completed job duration for the censored spells. The actual completed job duration is used for jobs for which the end is observed. This procedure has the advantage of using all available information on duration.

A. *The Jobs Sample*

All of the subsequent analysis is performed using data for male household heads aged 18–60 who participated in the *PSID*.[11] We used only observations from the random national sample portion of the *PSID* (the so-called Survey Research Center or SRC subsample). Persons who were retired, permanently disabled, self-employed, employed by the government, or residents of Alaska or Hawaii were excluded from the sample. Because we were concerned that different processes might govern tenure attainment and earnings in the union sector than in the nonunion sector, we also excluded observations on unionized jobs.[12] We were also concerned about differences across occupations in the processes determining job duration and earnings. In what follows, we therefore focus our discussion on results for two occupational subgroups. The first is a subset of white-collar occupations including nonunion professional, technical, and managerial employees.[13] The second subsample is comprised of nonunion blue-collar employees. In each year from 1968 through 1981 in which those individuals satisfying our selection criteria were household heads, information was available on number of years they had

ditional instrument yield identical estimates of the underlying parameters.

[11]Altonji kindly provided us with an extract containing the variables which we used in performing our analyses. The procedures followed in creating this extract are described in detail in an appendix to Altonji and Shakotko. In order to delete non-SRC subsample observations, we added information on whether a given individual was part of the SRC subsample or the nonrandom continuation subsample from the *Survey of Economic Opportunity*. We also smoothed the tenure variable in instances where a given individual had been assigned the midpoint value of a tenure interval. If the individual on a given job changed tenure intervals in succeeding years, we computed a smooth tenure variable forward and backward from the change point. If all observations for the individual on a given job were in the same tenure interval, we computed a smooth tenure variable forward and backward from the middle observed year on the job assuming that tenure in that year was equal to the midpoint of the interval.

[12]In some years, unionization refers to union membership, and in other years, to working on a job covered by a collective bargaining agreement. Where both measures were available, collective bargaining coverage was used. Observations on jobs for which the worker changed union status during the course of the job do not appear in the sample. In 371 jobs, workers were coded nonunion in some years and union in others. If 1) at least two-thirds of the observed years on one of these jobs were coded nonunion, 2) there were no runs of three or more years coded union, and 3) the first and last years observed on the job were coded nonunion, then the entire job was considered a nonunion job and was included in our sample.

[13]Excluded from the analysis were the approximately 16 percent of all nonunion jobs that were clerical and sales jobs. This is too few to support separate job duration models for these groups. At the same time, these jobs seemed likely to differ significantly from other white-collar jobs.

TABLE 1—SELECTED CHARACTERISTICS OF JOBS SAMPLES FOR OCCUPATIONAL SUBGROUPS[a]

	Managerial and Professional Nonunion			Blue Collar Nonunion		
	All	Complete	Censored	All	Complete	Censored
Proportion with Years of Tenure at Last Date Job Observed in Range:						
$T \leq 1$	.280	.483	.147	.483	.639	.294
$1 < T \leq 3$	.235	.251	.224	.215	.216	.213
$3 < T \leq 10$	.257	.205	.291	.109	.115	.268
$T > 10$	.228	.0614	.338	.119	.0310	.224
Mean of:[b]						
Years of Tenure at Last Date Job Observed	6.76 [8.05]	2.96 [4.03]	9.25 [9.00]	4.07 [6.43]	1.9 [3.2]	6.7 [8.1]
Years of Pre-Job Experience	9.76 [8.09]	10.00 [7.76]	9.61 [8.30]	10.68 [9.49]	9.8 [9.0]	11.8 [10.]
(Years Pre-Job Experience)^2	160.8 [253.7]	160.2 [230.1]	161.1 [268.1]	204.2 [348.6]	175.7 [325.1]	238.5 [372.0]
Years of Education	14.6 [2.09]	14.5 [2.1]	14.7 [2.1]	11.3 [2.43]	11.4 [2.3]	11.3 [2.6]
Proportion:						
Nonwhite	.0416	.0537	.0337	.138	.139	.137
Married	.862	.854	.867	.845	.823	.872
Disabled	.0528	.0563	.0505	.0932	.0761	.114
Managerial	.483	.517	.461	–	–	–
Prof., Tech.	.517	.483	.539	–	–	–
Foreman, Craft	–	–	–	.439	.421	.461
Oper., Labor	–	–	–	.561	.579	.539
No. of Observations	985	391	594	1417	775	642

[a]Except for tenure and years of previous experience, all variables are reported as of the first year the job was observed. Previous experience was computed as the difference between reported experience in the first year the job was observed and seniority at that point.
[b]Standard deviations are shown in brackets.

held their current job, number of years they had worked prior to taking the current job, years of education, race, marital status, disability status, occupation, industry, region, and earnings.

There are 985 jobs held by 706 individuals represented in the nonunion white-collar subsample, and 1417 jobs held by 831 individuals represented in the nonunion blue-collar sample. Our concern at this point is with ascertaining how long each of these jobs ultimately lasted.

Various characteristics of the jobs in each of the samples are reported in Table 1. Variables that can change over time in an unpredictable fashion (for example, marital status, occupation) are assumed constant and measured at the first point the job is observed in the sample. Any jobs for which there are some blue-collar years and some white-collar years appear in both occupational subsamples. The last observed seniority on a job is always considered to be the seniority at the last date the person is observed with an employer, whether or not there has been a change of occupation during the course of the job. There were 87 cases in which an individual reported moving from blue-collar status to white-collar status, and 83 cases in which an individual reported moving from white-collar status to blue-collar status while a job was in progress.[14]

We observe the actual completed duration for 391 of 985 jobs in the white-collar sam-

[14]It is likely that these numbers overstate the true number of changes since there are undoubtedly some errors in classification that produce spurious movements between the two broad occupational groups.

ple and for 775 of 1417 jobs in the blue-collar sample. Not surprisingly, a large proportion of the completed jobs are relatively short. However, in both samples, there are a sizable number of completed jobs lasting 3 to 10 years and over 10 years. Longer jobs are more common among the still-in-progress jobs.

B. *Specification and Estimation of the Job Duration Model*

The proportional hazard Weibull specification serves as the basis of the estimation reported here. In that specification, the probability that a job has completed duration (D) greater than or equal to T is

$$(17)\qquad Pr(D \geq T) = \exp[-\lambda T^{\tau}],$$

where τ is a positive parameter. The proportional hazard assumption implies that

$$(18)\qquad \lambda = e^{-Z\Gamma},$$

where Z is a vector of observable individual characteristics hypothesized to affect job duration and Γ is a vector of parameters.

If the parameters of this distribution are estimated, there is some ambiguity in the interpretation of the estimate of τ. The obvious interpretation is that the estimated value of τ indicates "true" duration dependence in the hazard of a job ending. An alternative interpretation is that the estimate of τ is biased downward by unmeasured heterogeneity in match quality. For the purposes of this study, we are interested only in an accurate estimate of completed duration, and we proceed with the simple Weibull specification.[15]

The contribution to the likelihood function made by a completed job is the probability-density that the job lasted *exactly* S_f years given that the job lasted at least S_0 years.[16] Given a Weibull distribution for duration, this is

$$(19)\qquad Pr(D = S_f | D \rangle S_0) = \lambda\tau S_f^{\tau-1} \exp\left[-\lambda\left(S_f^{\tau} - S_0^{\tau}\right)\right].$$

Similarly, the contribution to the likelihood function made by a job with a censored duration is the probability that the job lasted *more than* S_f years given that the job lasted at least S_0 years. This is

$$(20)\qquad Pr(D > S_f | D \rangle S_0) = \exp\left[-\lambda\left(S_f^{\tau} - S_0^{\tau}\right)\right].$$

The log-likelihood function is formed from these probabilities as

$$(21)\qquad \ln(L) = \sum_j \left\{C_j \ln Pr(D_j > S_{fj} | D_j \rangle S_{0j}) + (1 - C_j) \ln Pr(D_j = S_{fj} | D_j \rangle S_{0j})\right\},$$

where j indexes jobs and C_j is an indicator variable that equals one if the completed job duration is censored (i.e., the job does not end during the sample period) and equals zero otherwise (i.e., the completed job duration is observed).[17]

Table 2 contains estimates of the Weibull job duration model estimated over the subsamples of 985 white-collar jobs and 1417 blue-collar jobs, respectively. These estimates were derived by maximizing the likelihood function defined above with respect to the parameters Γ and τ.[18] In interpreting the

[15] See Tony Lancaster (1979) for a parametric approach to the problem of estimating unmeasured heterogeneity in a Weibull model of unemployment duration. James Heckman and Burton Singer (1984) present a nonparametric approach to estimating duration models with unmeasured heterogeneity.

[16] It is important to condition on the length of the job as of the date it is first observed because the sampling scheme is such that jobs will not be observed unless they last long enough to make it to the start of the sample period.

[17] Note that this specification of the likelihood function assumes that unmeasured factors affecting completed job durations are independent across spells. However, within each sample, there are multiple observations on job durations for some individuals. Given the nonlinear nature of the model, an appropriate tractable procedure for accounting for the induced correlation is not obvious.

[18] The algorithm described by Ernst Berndt et al. (1974) was used to find the maximum.

TABLE 2—SELECTED COEFFICIENTS FROM FINAL TENURE MODELS[a]

	Managerial and Prof. Nonunion (1)	Blue Collar Nonunion (2)
Γ (Inverse Baseline Hazard, $\lambda = e^{-z\Gamma}$)		
Years of Experience	−.0288	.0611
	(.0204)	(.0010)
(Years of Experience)^2	.00123	−.00113
	(.00071)	(.00027)
Years of Education	.0699	.0243
	(.0244)	(.0136)
Nonwhite (yes = 1)	−.412	−.0387
	(.224)	(.0878)
Married (yes = 1)	.270	.464
	(.135)	(.073)
Disabled (yes = 1)	.0768	.180
	(.218)	(.113)
Manager (yes = 1)	−.0764	–
	(.1121)	
Foreman, Craftworker (yes = 1)	–	.0656
		(.0623)
"Duration" Parameter		
τ	.380	.394
	(.028)	(.017)
Log-Likelihood	−900.4	−1097.9
Sample Size	985	1417

[a]These coefficient estimates are from a Weibull proportional hazards model implemented using the jobs samples described in Table 1. All explanatory variables are reported as of the start of the job. Professional/technical employees are the omitted occupational group in the col. 1 model and operatives/laborers are the omitted occupational group in the col. 2 model. The numbers shown in parentheses are asymptotic standard errors.

estimates of the determinants of the baseline hazard (λ), recall that the hazard rate was specified such that $\lambda = e^{-Z\Gamma}$. Thus, an increase in a variable with a positive coefficient reduces λ and increases the expected duration of the job. The hypothesis that the models of completed job duration for the two occupational groups are the same is strongly rejected.[19] The marginal effect of pre-job experience on job duration for white-collar workers is never statistically significant at the .05 level, while among blue-collar workers, having more pre-job experience has a significant positive association with completed job duration. The estimates also suggest that education has a stronger positive relationship with job duration in white-collar occupations than in blue-collar occupations.

C. *Prediction of Job Duration for Incomplete Jobs*

We used the parameter estimates from the appropriate column of Table 2 to predict the expected completed job duration of each of the incomplete jobs in each sample. This expectation is computed conditionally on the job lasting longer than the last observed seniority. Note that the job duration model we have estimated is based on data for the preretirement period. It will capture the net effects of quit and layoff processes on job duration, but it will not capture the effect of the competing retirement process which comes into play for older workers. If we predicted job durations without taking retirement into account, some would be implausibly long. We therefore assume that all jobs that are in progress when the worker reaches age 65 end at that point. For an individual/job match with observable characteristics Z that has lasted S_f years as of the last date we observe it, the conditional expected completed job duration is

$$(22)\quad \hat{E}(D|D\rangle S_f) = \frac{1}{Pr(D>S_f)}\int_{S_f}^{S_{65}} \lambda\tau t^{\tau} e^{-\lambda t^{\tau}} dt + \frac{Pr(D>S_{65})}{Pr(D>S_f)} * S_{65},$$

where S_{65} represents the seniority attained if a match lasts until the worker turns 65,

$$(23)\quad Pr(D\rangle S_f) = \exp[-\lambda S_f^{\tau}],$$

$$Pr(D \geq S_{65}) = \exp[-\lambda S_{65}^{\tau}],$$

and $$\lambda = e^{-Z\Gamma}.$$

[19]The hypothesis that the parameters of the models for the two subgroups are identical except for a constant shift and the occupation dummies in $Z\Gamma$ can be rejected at any reasonable level of significance. The test statistic, distributed as χ^2 with 22 degrees of freedom, is 52.3 (p-value < .001). The independence assumption of this test is not strictly satisfied, since the two samples contain some jobs in common.

TABLE 3—DISTRIBUTION OF COMPLETED JOB DURATIONS[a]

Proportion with Completed Job Duration in Range:	Managerial and Professional Nonunion			Blue Collar Nonunion		
	All	Complete	Censored	All	Complete	Censored
$D \leq 1$	.193	.483	.00168	.354	.639	.0109
$1 < D \leq 3$	.105	.251	.00842	.155	.216	.0826
$3 < D \leq 10$	.182	.205	.167	.253	.115	.419
$D > 10$	.521	.0614	.823	.238	.0310	.488
No. of Observations	985	391	594	1417	775	642

[a]For the completed job subsample, the distribution of actual completed job duration is shown. For the censored job subsample, the distribution of predicted completed job duration is reported.

With an appropriate change of variables, this expected duration can be computed numerically using incomplete *gamma* functions. We also use the square of completed job duration in the earnings function estimation. For incomplete jobs, this is estimated analogously to completed job duration.

As noted earlier, actual job duration was observed for a substantial fraction of the jobs in both samples, and this measure was used in these cases. For the jobs with censored duration, the predicted values were used. Table 3 contains the breakdown of the completed durations of the jobs in the two occupational samples. As expected, the censored jobs are generally longer than the completed jobs.

III. Is Seniority Half Completed Job Duration?

Samples of individual-year observations from the two *PSID* jobs samples just discussed are used to estimate the earnings functions which constitute the core of our analysis. Recall that the samples consist only of nonunion male heads of households. There are 3493 individual-year observations on the 706 workers in the 985 white-collar jobs, and 3554 individual-year observations on the 831 workers in the 1417 blue-collar jobs. The means and standard deviations of the central variables for each of the samples are contained in the first column of Tables 4a and 4b.

Two alternative approaches to removing the bias in the estimated return to seniority have been suggested: 1) using the residual from a regression of seniority on completed duration as an instrument for seniority, and 2) including completed duration as a regressor in the earnings function. Both approaches started from the recognition that a true random sample of years from jobs would have the property that, on average, seniority equals one-half completed duration. This is equivalent to the hypothesis that in a regression of seniority on completed duration the constant term is zero and the coefficient on completed duration is .5.

In order to investigate this directly, we regressed seniority on the completed duration for the two occupational subsamples. The results are

$$(24)\quad \underset{}{S_{ijt}} = \underset{(.18)}{-2.24} + \underset{(.007)}{.534} \cdot D_{ij} \qquad R^2 = .61 \quad \text{(white collar)},$$

$$S_{ijt} = \underset{(.097)}{-1.31} + \underset{(.005)}{.550} \cdot D_{ij} \qquad R^2 = .75 \quad \text{(blue collar)},$$

and the numbers in parentheses are OLS standard errors.[20] Clearly, the hypothesis

[20]Given that D_{ij} is a predicted value for the observations from censored jobs, the OLS standard errors are not appropriate. However, the inferences are unchanged by use of (computationally tedious) standard errors that account for the fact that D_{ij} is predicted as well as general heteroskedasticity of the form analyzed by Halbert White (1980). See Whitney Newey (1984) for computational details on the correct standard errors.

TABLE 4a—SELECTED COEFFICIENTS FROM ln (AVERAGE HOURLY EARNINGS) MODELS MANAGERIAL AND PROFESSIONAL NONUNION SAMPLE[a]

	Mean [s.d.]	OLS (1)	IV (2)	OLS (3)	OLS (4)
Years of Experience	18.14 [10.08]	.0349 (.0027)	.0392 (.0058)	.0288 (.0027)	.0263 (.0031)
(Years of Experience)^2	430.77 [407.84]	−.00062 (.00006)	−.00077 (.00014)	−.00048 (.00007)	−.00043 (.00007)
Years of Current Seniority	8.88 [8.34]	.0106 (.0011)	.00585 (.00128)	.00548 (.00178)	.00520 (.00256)
E(Completed Job Duration)	20.83 [12.18]	–	–	.0198 (.0024)	.0265 (.0050)
$\{E$(Completed Job Duration)$\}^2$	631.55 [505.56]	–	–	−.00035 (.00006)	−.00059 (.00016)
E(Job Duration) ×[=1 if 3 < Seniority ≤ 10]	6.02 [10.46]	–	–	–	−.00094 (.00432)
$\{E$(Job Duration)$\}^2$ ×[=1 if 3 < Seniority ≤ 10]	165.4 [325.3]	–	–	–	.00009 (.00015)
E(Job Duration) ×[=1 if Seniority > 10]	11.09 [15.78]	–	–	–	−.00798 (.00455)
$\{E$(Job Duration)$\}^2$ ×[=1 if Seniority > 10]	380.1 [572.9]	–	–	–	.00030 (.00015)
R^2	–	.3696	.3575	.3871	.3883

[a]All models also include controls for education, race, marital status, disability, occupation, industry, region, and year. E (Completed Job Duration) is computed using the estimates in col. 1 of Table 2. The numbers shown in parentheses are standard errors. Sample size = 3493.

TABLE 4b—SELECTED COEFFICIENTS FROM ln (AVERAGE HOURLY EARNINGS) MODELS BLUE-COLLAR NONUNION SAMPLE[a]

	Mean [s.d.]	OLS (1)	IV (2)	OLS (3)	OLS (4)
Years of Experience	17.34 [11.14]	.0205 (.0024)	.0173 (.0040)	.0117 (.0026)	.0120 (.0026)
(Years of Experience)^2	424.70 [470.81]	−.00045 (.00006)	−.00042 (.00009)	−.00026 (.00006)	−.00028 (.00006)
Years of Current Seniority	6.31 [7.46]	.0142 (.0011)	.00290 (.00172)	.00241 (.00213)	−.00054 (.00302)
E(Completed Job Duration)	13.86 [11.75]	–	–	.0154 (.0021)	.0381 (.0057)
$\{E$(Completed Job Duration)$\}^2$	362.44 [444.45]	–	–	−.00014 (.00006)	−.00104 (.00024)
E(Job Duration) ×[=1 if 3 < Seniority ≤ 10]	4.57 [8.27]	–	–	–	−.00592 (.00538)
$\{E$(Job Duration)$\}^2$ ×[=1 if 3 < Seniority ≤ 10]	102.13 [211.7]	–	–	–	.00031 (.00024)
E(Job Duration) ×[=1 if Seniority > 10]	6.45 [12.90]	–	–	–	−.0241 (.0055)
$\{E$(Job Duration)$\}^2$ ×[=1 if Seniority > 10]	215.1 [461.9]	–	–	–	.00103 (.00024)
R^2	–	.3878	.3513	.4041	.4098

[a]All models also include the controls listed in Table 4a, fn. a. E (Completed Job Duration) is computed using col. 2, Table 2. Standard errors are shown in parentheses. Sample size = 3554.

that the coefficient on completed duration is .5 can be rejected in both samples at conventional levels of significance. This means that completed duration is not orthogonal to the deviation of seniority from one-half completed duration so that this deviation is not a valid instrument for seniority. However, the coefficient of completed duration is not far from one-half, which suggests that the general approach is valid.

On the basis of these results, the IV estimation of the earnings function proceeds using the residuals from the regression of seniority on completed duration to instrument seniority. We also present OLS estimates of earnings functions augmented with our measure of completed duration.

IV. Earnings Function Estimates

Consider an earnings function for individual i in job j in year t of the form:

$$(25)\quad \ln(W_{ijt}) = \theta_0 + \theta_1 S_{ijt} + \theta_2 E_{ijt} + \theta_3 E^2_{ijt} + X_{ijt}\Omega + \varepsilon_{ijt},$$

where $\ln(W_{ijt})$ is the logarithm of real average hourly earnings, S_{ijt} is seniority, E_{ijt} is total experience, X_{ijt} is a vector of other individual characteristics, and ε_{ijt} represents unmeasured factors affecting earnings.[21] The coefficient θ_1 is the net return to seniority and corresponds to $\beta_1-(\beta_2+\alpha)$ in the model of Section I. The coefficient θ_2 is the return to general labor market experience and corresponds to $\beta_2+\alpha$ in the model of Section I.

Tables 4a and 4b contain estimates of earnings functions for the samples of professional, technical, and managerial workers and of blue-collar workers, respectively. In addition to the regressors shown, all models include controls for education, race, marital status, disability, occupation, industry, region, and year. The standard errors presented are the "simple" standard errors computed from $\hat{\sigma}^2(X'X)^{-1}$ for the OLS models and from $\hat{\sigma}^2(Z'X)^{-1}$ for the IV models.[22]

In both tables, column 1 contains a standard OLS earnings equation that neither instruments for seniority nor includes completed job duration as a regressor. These estimates suggest that there are sizable returns both to general labor market experience and to seniority with a particular employer for workers in both occupational groups. The estimated net return to seniority is on the order of 1 to 1.5 percentage points per year.

The column 2 models were estimated by IV using pre-job experience, the square of pre-job experience, and the residual from the regression of seniority on completed job duration as instruments for total experience, the square of total experience, and seniority. The IV coefficient estimates provide consistent estimates of the return to general labor market experience and the net return to seniority. While instrumenting has relatively little effect on the estimated return to experience, the estimated net return to seniority falls substantially. The return for white-collar workers falls from 1.1 to 0.6 percent per year. The return for blue-collar workers falls from 1.4 to 0.3 percent per year; moreover, the corrected estimate is not significantly different from zero at conventional levels. This suggests that most of the cross-sectional correlation between earnings and seniority controlling for experience reflects the influence of omitted variables.

The coefficients in column 3 were estimated using ordinary least squares, but with completed job duration and its square added to the list of explanatory variables. The first

[21] We have estimated all of the models in this section with pre-job experience and its square in place of total experience and its square, and the qualitative conclusions emerging from the analysis do not change.

[22] These standard errors are not strictly appropriate for the estimation here because they do not account for the fact that the measure of completed job duration is predicted for the observations on censored jobs. Standard errors that are corrected both for this fact and for general heteroskedasticity (see Newey and White) were computed for a number of specifications. These were uniformly very close to the simple standard errors, and in no case was any inference resulting from a statistical test changed.

thing to note about these augmented OLS models is that the estimated returns to seniority are virtually identical to those obtained using the IV approach. This confirms that a substantial portion of the usual cross-sectional return to seniority is due to an omitted worker, job, and/or match quality measure. The virtual equality of the results using the augmented OLS approach and the IV approach also suggests that the bias in the augmented OLS estimates that we discussed in Section I is not a problem in practice.[23]

Perhaps the most interesting aspect of the augmented OLS estimates is that there is a *very* strong positive association between completed duration and earnings in both occupational groups. Consider two otherwise equivalent workers, one of whom holds a job that will eventually last 20 years and the other of whom holds a sequence of two 10-year jobs. In a white-collar occupation, the worker in the single 20-year job is estimated to earn 9.3 percent (standard error = 0.8 percent) more *in each year* than the worker in the sequence of 10-year jobs. In a blue-collar occupation, the worker in the single 20-year job is estimated to earn 11.2 percent (standard error = 1.0 percent) more *in each year* than the worker in the sequence of 10-year jobs.

The finding that workers in longer jobs earn more in every year on the job than workers in shorter jobs is verified by the results in column 4. These results are based on specifications that allow the effect of completed duration on earnings to differ by seniority level. Specifically, completed duration and its square are included in the regression along with interactions of these two variables with two dummy variables indicating seniority of 3 to 10 years and seniority greater than 10 years. The hypothesis that these additional four variables have zero coefficients cannot be rejected for the white-collar sample, but can be rejected for the blue-collar sample.[24] Closer examination of the estimated parameters reveals that, consistent with the results of the formal test, the four interaction terms have inconsequential coefficients in the white-collar sample. For blue-collar workers, the wage advantage of long jobs falls with seniority after starting at a much higher level than suggested by the results in column 3 of Table 4b. Thus, there is still a positive wage advantage to being in longer blue-collar jobs at all levels of seniority. It simply is not uniform in magnitude.

Overall, the results in this section strongly confirm our expectations. Rather small estimates of the return to seniority are found using either the IV or the augmented OLS approach relative to the standard cross-sectional OLS regression. The corrected estimate of the net return to seniority is 0.5 to 0.6 percent per year for white-collar workers and a statistically insignificant 0.2 to 0.3 percent per year for blue-collar workers. In addition, our results provide direct evidence that workers in longer jobs earn substantially more throughout the job than workers in shorter jobs.

V. Some Issues of Specification and Estimation

Probably the most significant potential problem with the analysis just described is that the key job duration variable had to be estimated for jobs whose end was not observed. This introduces three conceptually distinct sources of measurement error that could affect our estimates.

The first source of measurement error is that the expected value is used in place of the actual realization of completed job duration. This is not a problem so long as the correct parameters and the correct model have been used in computing expected job duration. In this case, the measurement error

[23] Recall that the theoretical discussion implied that the IV estimate of the net return to seniority is a consistent estimate of $\beta_1 - [\beta_2 - \alpha]$ while the augmented OLS approach yields an upward biased estimate. In fact, we find that the augmented OLS estimate of the net return to seniority is slightly, though not significantly, *smaller* than the IV estimate.

[24] The test statistic for the white-collar sample is 1.69 and that for the blue-collar sample is 8.48, both distributed as F with 4 and approximately 3500 degrees of freedom. The critical value of this distribution is 2.37 at the 5 percent level and 3.32 at the 1 percent level.

is exactly the deviation between expected and actual job duration. This is uncorrelated with expected job duration (the included regressor) by construction. Thus, there is no bias from this source in our estimated earnings function coefficients.

The second source of measurement error is that the parameters of the job duration model are only estimates, so that the predictions of expected job duration are themselves subject to error. However, it can be shown that in large samples, the estimation error in the parameters of the job duration model is of small enough order that coefficient estimates in equations which use the derived measure of duration as an explanatory variable are consistent.

The third source of measurement error is that the job duration model may be misspecified. If misspecification results in random errors in expected completed duration then classical measurement error is introduced. In the context of the augmented OLS estimates, this measurement error will tend to 1) bias the coefficient on completed duration toward zero, and 2) result in an estimated return to seniority that is higher than it would be in the absence of measurement error. The estimated return to seniority would, however, still be useful as an upper bound. If misspecification results in systematic measurement error, the coefficient on completed duration and the estimated return to seniority will be biased in an unknown way.

Given the finding of a large positive return to completed duration and the sharp reduction in the return to seniority using the augmented OLS estimates, it is unlikely that random measurement error is a serious problem. In any case, since the effects of random measurement error are predictable, our findings provide bounds on the "true" effects. We can do nothing about the potential of systematic measurement error except to note that our results are conditional on the Weibull specification of completed job duration.[25]

The appropriateness of the method used to impute completed durations for the censored jobs clearly merits careful investigation. One obvious question is whether the particular function of the explanatory variables and last-observed seniority that we use in creating our measure of completed duration is appropriate or whether it is contributing to the fit of the model simply because it incorporates the information on last-observed seniority. One way to investigate this is to reestimate the augmented OLS model including last-observed seniority and its square directly as regressors. If our measure of completed job duration has a significant effect on earnings even after controlling for last-observed seniority, then we have added support for our measure.

The first columns of Tables 5a and 5b contain estimates of earnings functions that include measures of completed job duration and its square, but not last-observed seniority (repeated here for comparison purposes). The second columns of these tables contain estimates of earnings functions that include last-observed seniority and its square, but no completed duration measure. When entered separately, both completed job duration and last-observed seniority contribute significantly to the fit of the model. Do these relationships hold up if we control for both simultaneously? The third column of Tables 5a and 5b contain estimates of earnings functions for the two occupational groups that include measures of both completed duration and last observed seniority. After controlling for completed duration, last-observed seniority is not a significant determinant of earnings. However, even after controlling for last observed seniority, completed duration adds significantly to the models' explanatory power.[26] Thus, our

[25]While the first two sources of measurement error do not induce inconsistency in the earnings function parameter estimates, they do affect the estimates of the standard errors of the coefficients. As discussed in fn. 22, appropriate standard errors that are corrected for the effects of these errors and for general heteroskedasticity are almost identical to the usual standard errors.

[26]This is verified by F-tests of the general specification in col. 3 against the restricted specifications in cols. 1 and 2. The test statistics for the hypothesis that

TABLE 5a—SELECTED COEFFICIENTS FROM ALTERNATIVE ln (AVERAGE HOURLY EARNINGS) MODELS MANAGERIAL AND PROFESSIONAL NONUNION SAMPLE[a]

	Mean [s.d.]	OLS (1)	OLS (2)	OLS (3)	OLS (4)
Years of	18.14	.0288	.0272	.0280	.0288
Experience	[10.08]	(.0027)	(.00280)	(.0028)	(.0027)
(Years of	430.77	−.00048	−.00042	−.00046	−.00051
Experience)^2	[407.84]	(.00007)	(.00007)	(.00007)	(.00007)
Years of Current	8.88	.00548	−.00472	.00186	.00865
Seniority	[8.34]	(.00178)	(.00311)	(.00345)	(.00244)
E(Completed Job	20.83	.0198	–	.0167	.0220
Duration)	[12.18]	(.0024)		(.0041)	(.0025)
{E(Completed Job	631.55	−.00035	–	−.00030	−.00044
Duration)}^{2}	[505.56]	(.00006)		(.00009)	(.00007)
Years Last Observed	12.08	–	.0275	.00682	–
Seniority	[9.52]		(.0031)	(.00567)	
{Years Last Observed	236.8	–	−.00043	−.00007	–
Seniority}^{2}	[304.9]		(.00008)	(.00012)	
E(Job Duration)	12.15	–	–	–	.00067
×[=1 if Uncensored]	[34.45]				(.00592)
{E(Job Duration)}^{2} ×	133.41	–	–	–	−.00030
×[=1 if Uncensored]	[629.26]				(.00028)
R^2	–	.3871	.3841	.3874	.3885

[a]See Table 4a, fn. a. Sample size = 3493.

TABLE 5b—SELECTED COEFFICIENTS FROM ALTERNATIVE ln (AVERAGE HOURLY EARNINGS) MODELS BLUE-COLLAR NONUNION SAMPLE[a]

	Mean [s.d.]	OLS (1)	OLS (2)	OLS (3)	OLS (4)
Years of	17.34	.0117	.0135	.0118	.0119
Experience	[11.14]	(.0026)	(.00256)	(.00259)	(.0027)
(Years of	424.70	−.00026	−.00028	−.00026	−.00027
Experience)^2	[470.81]	(.00006)	(.00006)	(.00006)	(.00006)
Years of Current	6.31	.00241	−.00304	.00375	.00412
Seniority	[7.46]	(.00213)	(.00322)	(.00352)	(.00307)
E(Completed Job	13.86	.0154	–	.0175	.0167
Duration)	[11.75]	(.0021)		(.0052)	(.0022)
{E(Completed Job	362.44	−.00014	–	−.00017	−.00020
Duration)}^{2}	[444.45]	(.00006)		(.00012)	(.00008)
Years Last Observed	8.80	–	.0255	−.00377	–
Seniority	[8.82]		(.0031)	(.00730)	
{Years Last Observed	155.2	–	−.00031	.00005	–
Seniority}^{2}	[263.8]		(.00008)	(.00017)	
E(Job Duration)	12.92	–	–	–	.00556
×[=1 if Uncensored]	[33.92]				(.00627)
{E(Job Duration)}^{2}	13.77	–	–	–	−.00053
×[=1 if Uncensored]	[60.21]				(.00030)
R^2	–	.4041	.4001	.4042	.4051

[a]See Table 4b, fn. a. Sample size = 3554.

measure of completed duration contains information on earnings well beyond the information contained directly in the variables that are used in its computation, including the last-observed seniority.

Perhaps the most important question concerning our completed job duration variable is whether the relationship between completed job duration and earnings differs between the observations that come from jobs where actual completed durations are observed and from jobs where completed durations are predicted. To answer this question, we reestimated the augmented OLS earnings functions with two additional variables: 1) the interaction between completed job duration and a dummy variable that equals one if the job duration is uncensored and equals zero otherwise; and 2) the interaction between the square of completed job duration and the same dummy variable. This allows completed duration and its square to have different effects where they are actually observed (uncensored jobs) and where they are predicted (censored jobs). These estimates are contained in column 4 of Tables 5a and 5b.

For white-collar workers, the hypothesis that the effects are the same (that the two new variables have zero coefficients) can be rejected at the 5 percent level, but not at the 1 percent level. However, the interaction terms have rather small coefficients relative to the coefficients on the basic variables. For blue-collar workers, the hypothesis that the effects are the same cannot be rejected at either the 5 or 1 percent level, and the point estimates of the interaction terms' coefficients are insubstantial.[27] In sum, observed completed duration in uncensored jobs and our estimate of completed duration in censored jobs have almost identical relationships with earnings.

The last potential issue we consider here is related to the fact that, because we use pooled time-series cross-section data to estimate our earnings functions, there are repeated observations on particular individuals. If the earnings function residuals are correlated across observations within individuals, our standard errors are likely to be understated. We are reluctant to present estimates of a fixed-effect model because all of the variation in the measure of completed duration in such a model comes from those workers who change jobs within the sample period. These job changes will be dominated by short jobs which are not representative of the sample of jobs as a whole.[28] A *very* conservative alternate approach to this problem is to reestimate the key specifications of the model on a single-year cross section. If the inferences from such a specification are similar to those from the pooled model, then we can have more confidence in the overall validity of the results. The estimation of a single-year cross section also has the advantage that it provides results directly comparable to much of the existing literature on earnings functions.

In order to carry out this analysis, 1973 was selected as a representative year, and the analyses of Tables 4a and 4b were repeated on 1973 cross sections of 244 white-collar workers and 240 blue-collar workers. The single-year estimates are contained in Tables 6a and 6b, and two things are clear from a comparison of these results with the results in Tables 4a and 4b. First, as we expected, the results are less precisely estimated due to the much smaller sample sizes. Second, the results are very similar in character those derived using the pooled samples. The "standard" OLS results in column 1 of Tables 6a and 6b show statistically significant net returns to seniority. The IV estimates of the return to seniority are much

last-observed seniority adds explanatory power to the model are 0.845 for white-collar workers and 0.295 for blue-collar workers. The test statistics for the hypothesis that completed job duration adds explanatory power are 9.30 for white-collar workers and 12.1 for blue-collar workers. All the test statistics are distributed as F with 2 and approximately 3500 degrees of freedom. The critical value of this distribution is 4.61 at the 1 percent level of significance.

[27]The test statistics are 3.95 for the white-collar sample and 2.95 for the blue-collar sample, both distributed as F with 2 and approximately 3500 degrees of freedom. The critical value of this distribution is 3.00 at the 5 percent level and 4.61 at the 1 percent level.

[28]Note that year effects are removed in all specifications through the use of a complete set of year dummies.

TABLE 6a—SELECTED COEFFICIENTS FROM 1973 CROSS SECTION ln (AVERAGE HOURLY EARNINGS) MODELS MANAGERIAL AND PROFESSIONAL NONUNION SAMPLE[a]

	Mean [s.d.]	OLS (1)	IV (2)	OLS (3)	OLS (4)
Years of Experience	18.38 [10.12]	.0349 (.0104)	.0514 (.0199)	.0273 (.0106)	.0278 (.0126)
(Years of Experience)^2	440.06 [402.58]	−.00066 (.00025)	−.00111 (.00049)	−.00049 (.00025)	−.00050 (.00029)
Years of Current Seniority	9.00 [7.97]	.0135 (.0043)	.00653 (.00524)	.00721 (.00684)	.0109 (.0104)
E(Completed Job Duration)	20.99 [13.00]	–	–	.0230 (.0090)	.0432 (.0181)
$\{E$(Completed Job Duration)$\}^2$	642.94 [534.27]	–	–	−.00041 (.00023)	−.00095 (.00054)
E(Job Duration) $\times[=1$ if $3<$ Seniority $\leq 10]$	6.40 [10.92]	–	–	–	−.0180 (.0159)
$\{E$(Job Duration)$\}^2$ $\times[=1$ if $3<$ Seniority $\leq 10]$	174.22 [341.65]	–	–	–	.00044 (.00051)
E(Job Duration) $\times[=1$ if Seniority $>10]$	11.19 [16.01]	–	–	–	−.0231 (.0173)
$\{E$(Job Duration)$\}^2$ $\times[=1$ if Seniority $>10]$	385.70 [585.00]	–	–	–	.00060 (.00054)
R^2	–	.4318	.4105	.4566	.4640

[a]See Table 4a, fn. a. Sample size = 244.

TABLE 6b—SELECTED COEFFICIENTS FROM 1973 CROSS SECTION ln (AVERAGE HOURLY EARNINGS) MODELS BLUE COLLAR NONUNION SAMPLE[a]

	Mean [s.d.]	OLS (1)	IV (2)	OLS (3)	OLS (4)
Years of Experience	18.47 [11.32]	.0340 (.0100)	.0133 (.0149)	.0205 (.0099)	.0188 (.0099)
(Years of Experience)^2	468.83 [484.73]	−.00075 (.00020)	−.00036 (.00033)	−.00046 (.00022)	−.00046 (.00022)
Years of Current Seniority	6.50 [7.60]	.0102 (.0038)	−.00453 (.00646)	.00063 (.00840)	−.00781 (.0122)
E(Completed Job Duration)	14.57 [12.88]	–	–	.0194 (.0075)	.0502 (.0208)
$\{E$(Completed Job Duration)$\}^2$	402.81 [491.00]	–	–	−.00028 (.00024)	−.00161 (.00088)
E(Job Duration) $\times[=1$ if $3<$ Seniority $\leq 10]$	4.86 [8.95]	–	–	–	.00108 (.0199)
$\{E$(Job Duration)$\}^2$ $\times[=1$ if $3<$ Seniority $\leq 10]$	114.07 [237.2]	–	–	–	.00031 (.00086)
E(Job Duration) $\times[=1$ if Seniority $>10]$	7.05 [13.81]	–	–	–	−.0381 (.0203)
$\{E$(Job Duration)$\}^2$ $\times[=1$ if Seniority $>10]$	245.9 [508.6]	–	–	–	.00171 (.00089)
R^2	–	.4958	.4048	.5182	.5344

[a]See Table 4b, fn. a. Sample size = 240.

smaller than the OLS estimates, and they are *not* significantly different from zero in either occupational group. The augmented OLS estimates of the return to seniority, contained in column 3 of Tables 6a and 6b, are quite similar to those derived with the IV estimator. In addition, the augmented OLS estimates imply that workers in longer jobs earn significantly more throughout the job than workers in shorter jobs. We conclude that the general findings of the previous section are not simply due to an exaggerated precision that comes from ignoring the error component structure in a pooled sample.

VI. Concluding Comments

The basis for considering implicit contracts under which compensation is deferred from early until late in workers' time with their employers to be an important feature of the labor market has been the simple cross-sectional evidence that long seniority workers have higher wages, even taking their total labor-market experience into account. The evidence presented in this study seriously undermines the empirical foundations of this sort of implicit contract. Contrary both to the conventional wisdom and to our own prior expectations, there seems to be only a small average return to seniority in excess of the average return to general labor market experience. For the nonunion professional, technical, and managerial sample, the corrected estimates of the seniority coefficient suggest that the true return to seniority is approximately half a percent per year, rather than the approximately 1 percent per year suggested by the standard cross-section model. For the nonunion blue-collar sample, the corrected estimates yield a statistically insignificant return to seniority of approximately one-quarter of 1 percent per year, rather than the statistically significant 1.5 percent per year suggested by the standard model.[29]

This evidence does not imply that implicit contracts entailing the posting of a bond by workers through a deferral of compensation are never important. Indeed, they could be very important for some subgroups of workers and even a small return to seniority could translate into a substantial cumulative contribution to annual earnings over a period of time. It is also possible that parts of the total compensation package other than earnings, such as fringe benefits or other perquisites, are structured so as to reward longevity with a particular employer.[30] However, earnings deferral under implicit contracts appears to be a much less important factor in both white- and blue-collar labor markets than has generally been believed.

The other result of our study that is potentially very important for the understanding of the nature of the long-term employment relationship is that workers in long jobs earn substantially more throughout their jobs than do workers in short jobs. Whether this is due to individual differences, inter-job differences, or match-specific differences, this finding has important implications for the decisions of workers and employers as they affect investment in match-specific capital and the dynamics of the employment relationship. Note that our finding of a correlation between completed duration and earnings does *not* imply that it is the length of the job that *causes* earnings. It is likely that, at least to some extent, the higher earnings throughout the job provide an incentive for workers to remain on their job. Viewed in this light, our results are consistent with the currently popular efficiency wage models.

The finding of a strong positive relationship between job duration and earnings in both of the occupational subsamples provides strong evidence against another view of long-term employment relationships based on models of incomplete information where firms offer workers insurance regarding their unknown abilities. In particular, Milton Harris and Bengt Holmstrom (1982) argue

[29] Our findings regarding the returns to seniority are consistent with those obtained by Altonji and Shakotko.

[30] Freeman and Medoff provide evidence that the value of nonwage benefits such as vacations and pension plans rise with seniority. There may also be less tangible advantages that accrue with seniority.

that the positive return to experience found in most data sets could reflect such insurance contracts. A simple version of this model considers the case where a worker and all firms are initially uncertain about the worker's productivity, and where that uncertainty is reduced over time as the worker and all firms learn about the worker's productivity from the worker's stream of output. The optimal contract for the firm to offer the worker specifies an initial wage equal to the expected value of the worker's productivity minus an insurance premium, and the employer guarantees not to reduce the initial wage if the worker is revealed to be relatively unproductive. Since all new information about productivity is common knowledge, workers revealed to be relatively productive receive wage increases either from their original employer or by taking a new job with another employer. Workers revealed to have low productivity cannot duplicate their original wage elsewhere and for that reason are more likely to stay with their original employer. Thus, this simple insurance model predicts a *negative* correlation between job duration and earnings. Our results indicate that this correlation is strongly positive.

The recognition that the direction of causality between earnings and job duration is ambiguous highlights the difficulty of using our results to make definitive statements on the existence of particular types of labor contracts. Observed job durations are generated as the result of mobility decisions that are poorly understood. While we conclude tentatively that the evidence from earnings functions is not consistent with simple earnings deferral models of incentive contracts or with the simple model of insurance contracts, it is clear that further analysis of the joint determination of job duration and earnings is necessary for a full understanding of long-run employment relationships.

REFERENCES

Abraham, Katharine G. and Medoff, James L., "Length of Service and the Operation of Internal Labor Markets," *Proceedings of the Thirty-Fifth Annual Meeting of the Industrial Relations Research Association*, December 1982, 308–18.

Akerlof, George A. and Katz, Lawrence F., "Do Deferred Wages Dominate Involuntary Unemployment as a Worker Discipline Device?," NBER Working Paper No. 2025, September 1986.

Altonji, Joseph and Shakotko, Robert, "Do Wages Rise with Job Seniority?," *Review of Economic Studies*, forthcoming 1987.

Becker, Gary S., *Human Capital*, NBER, Ann Arbor: University Microfilms, 1964.

______ **and Stigler, George J.**, "Law Enforcement, Malfeasance and Compensation of Enforcers," *Journal of Legal Studies*, January 1974, *3*, 1–18.

Berndt, Ernst K. et al., "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 1974, *3/4*, 653–65.

Bulow, Jeremy and Summers, Lawrence, "A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynsian Unemployment," *Journal of Labor Economics*, July 1986, *4*, 376–415.

Burdett, Kenneth, "A Theory of Employee Job Search and Quit Rates," *American Economic Review*, March 1978, *68*, 212–20.

Freeman, Richard B. and Medoff, James L., *What Do Unions Do?*, New York: Basic Books, 1984.

Harris, Milton and Holmstrom, Bengt, "Ability, Performance and Wage Differentials," *Review of Economic Studies*, July 1982, *49*, 315–33.

Heckman, James and Singer, Burton, "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, March 1984, *52*, 271–320.

Jovanovic, Boyan, "Job Matching and the Theory of Turnover," *Journal of Political Economy*, October 1979, *87*, 972–90.

Katz, Lawrence F., "Efficiency Wage Theories: A Partial Evaluation," in Stanley Fischer, ed., *NBER Macroeconomics Annual 1986*, Cambridge: MIT Press, 235–76.

Lancaster, Tony, "Econometric Methods for the Duration of Unemployment," *Econometrica*, July 1979, *47*, 939–56.

Lazear, Edward, "Why Is There Mandatory Retirement?," *Journal of Political Econ-*

omy, December 1979, *87*, 1261–84.

Medoff, James L. and Abraham, Katharine G., "Experience, Performance and Earnings," *Quarterly Journal of Economics*, December 1980, *95*, 703–36.

______ and ______, "Are Those Paid More Really More Productive: The Case of Experience," *Journal of Human Resources*, Spring 1981, *16*, 186–216.

Mincer, Jacob, *Schooling, Experience and Earnings*, NBER *Studies in Human Behavior and Social Institutions*, No. 2, New York: Columbia University Press, 1974.

______ and Jovanovic, Boyan, "Labor Mobility and Wages," in Sherwin Rosen, ed., *Studies in Labor Markets*, NBER Universities-National Bureau Conference Series, No. 31, Chicago: University of Chicago Press, 1981, 21–64.

Mortensen, Dale T., "Specific Capital and Labor Turnover," *Bell Journal of Economics*, Autumn 1978, *9*, 572–86.

Newey, Whitney K., "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 1984, *14*, 201–06.

Parsons, Donald O., "Models of Labor Market Turnover: A Theoretical and Empirical Survey," in Ronald G. Ehrenberg, ed., *Research in Labor Economics*, Vol. 1, Greenwich: JAI Press, 1977, 185–223.

Shapiro, Carl and Stiglitz, Joseph, "Involuntary Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Stiglitz, Joseph, "Theories of Wage Rigidity," mimeo., Princeton University, 1984.

Topel, Robert, "Job Mobility, Search, and Earnings Growth: A Reinterpretation of Human Capital Earnings Functions," in Ronald G. Ehrenberg, ed,. *Research in Labor Economic*, Vol. 8, 1986, 199–223.

Viscusi, W. Kip, "Self-Selection, Learning-Induced Quits and the Optimal Wage Structure," *International Economic Review*, October 1980, *21*, 529–46.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48*, 817–38.

Yellen, Janet, "Efficiency Wage Models of Unemployment," *American Economic Review Proceedings*, May 1984, *74*, 200–08.

Savings of the Elderly and Desired Bequests

By MICHAEL D. HURD*

Cross-section data often show that the wealth of the elderly increases with age, suggesting that the life cycle hypothesis of consumption should include a bequest motive for saving. I propose a model of bequests, and a test for a bequest motive. Empirical findings are that in a ten-year panel data set, the elderly dissaved, in contradiction to most cross-section results. The test offers no support for a bequest motive.

Although the life cycle hypothesis of consumption has played a central role in theoretical and empirical work about consumption since it was proposed by Franco Modigliani and Richard Brumberg in 1954, many recent empirical studies have failed to support the hypothesis. In particular, cross-section data suggest that the wealth of the elderly increases with age, even at advanced ages. This result has been interpreted to support an extended life cycle hypothesis in which consumers derive utility from both consumption and bequests; that is, consumers have a bequest motive for saving. In this paper I propose and analyze a model of consumption and bequests, and specify a test for a bequest motive. Empirical results from panel data show that, in contradiction to most cross-section studies, the elderly dissave after retirement. According to my test, the data offer no support for a bequest motive.

Three kinds of studies have contributed to the doubt about the strict life cycle hypothesis: simulation and estimation of earnings and consumption paths, studies based on time-series methods, and micro data estimation of the age-wealth relationship. I briefly review the results of the first two and discuss the third in considerably more detail. See Mervyn King (1985) for a recent survey.

Studies that simulate the consumption and earnings paths of households (Betsey White, 1978, 1984; Michael Darby, 1979) or estimate the paths directly (Laurence Kotlikoff and Lawrence Summers, 1981) typically show that the aggregate of the present value of savings can account for only a small fraction of the capital stock that is held by households. Because the holdings of capital stock not generated by household saving must have been inherited, bequests must account for a large portion of the capital stock. These authors conclude that the strict life cycle hypothesis cannot be true for an important fraction of the population.

Other authors have used time-series methods to study consumption (Robert Hall, 1978, 1985; Marjorie Flavin, 1981; and Fumio Hayashi, 1982, 1985). The objective of these studies is to estimate the parameters of a stochastic difference equation for consumption. In this framework, the life cycle hypothesis makes the strong prediction that, given lagged consumption, the influence of wealth and income on current consumption will be zero. Often these studies reject the life cycle hypothesis at least as a hypothesis governing the behavior of all consumers (see Hall and Frederic Mishkin, 1982).

The relationship between wealth and age that is generally found in cross-section data is implausible according to the life cycle hypothesis: the elderly seem to accumulate wealth as they age even though the life cycle

*Department of Economics, SUNY, Stony Brook, NY 11794, and National Bureau of Economic Research. I am grateful to Thomas Prusa for excellent research assistance, and to Ken Hendricks for valuable comments and suggestions. The research reported herein was performed pursuant to grant no. 84ASPE138A from the U.S. Department of Health and Human Services. The opinions and conclusions expressed are solely my own and should not be construed as representing the opinions or policy of any agency of the U.S. government.

hypothesis implies they should decumulate (Thad Mirer, 1979; Paul Menchik and Martin David, 1983; and Mordecai Kurz, 1984). The typical finding in cross-section data is summarized by this quotation from Sheldon Danziger et al: "The elderly not only do not dissave to finance their consumption during retirement, they spend less on consumption goods and services (save significantly more) than the nonelderly at all levels of income. Moreover, the oldest of the elderly save the most at given levels of income" (1982, p. 210). The empirical finding that the elderly seem not to dissave has probably had the greatest effect in convincing economists that the strict life cycle hypothesis is not valid. The reasoning is that there is a maximum age to which people can live, and, without a bequest motive, people will want to consume all their wealth by that age. Yet, wealth seems to increase at any age. The conclusion is that there must be a bequest motive.

I believe there are fundamental difficulties in drawing such an inference from cross-section results. The wealthy tend to live longer than the poor; therefore, the wealth holdings of older people will be *ceteris paribus* higher than the wealth holdings of younger people. In addition, each cohort will have had a different lifetime income level, and rate of return on investments. Some adjustment, especially for lifetime income, must be made, or comparisons across age groups will be meaningless. In that the adjustment for each cohort cannot be estimated in the cross-section data, it has to be imposed; for example, it is often assumed to follow long-term trends such as growth rates of wages. This means that lifetime income at each age is adjusted by the long-term trend with the greatest ages having the greatest adjustment. Whether one adjusts observed income to estimate lifetime income (King and Louis Dicks-Mireaux, 1982), or adjusts wealth itself (Mirer), the age profile could slope up or down depending on the adjustment that is chosen. Thus, the adjustment itself, rather than the data, could determine the relationship between wealth and age. My final reason for not having much confidence in the cross-section studies is that it is very difficult in cross-section data to be certain that people have retired. Because some of the young elderly are still working, wealth will initially increase with age even after normal retirement age. It is certainly not inconsistent with the life cycle hypothesis that the wealth of workers increases with age.

Peter Diamond and Jerry Hausman (1984) find, using the *National Longitudinal Survey* of older men, that the elderly dissave after retirement. This data set is not well-suited for a study of the wealth of the elderly after they retire because by the end of the ten-year panel the ages of the sample range from 55 to 69. Even with a retirement age of 62, which is earlier than average, only half of the sample would be retired in the last year of the survey; therefore, wealth changes of retired people can only be observed for a few years, and, even then, most of the retired will be early retirees who may not be typical in their savings behavior. Furthermore, it is difficult to judge the results because no definition of wealth is given in their paper.

Douglas Bernheim (1987) studied wealth changes of retired individuals and couples from panel data. In his sample, wealth generally declined between 1969 and 1975, and between 1975 and 1979. This is good evidence that the elderly dissave; but it is not conclusive because Bernheim only observed two wealth changes, and because he used only a small fraction of the sample.[1]

In this paper I offer evidence on the empirical validity of the strict life cycle hypothesis against the life cycle hypothesis with bequests. I propose a model of bequests that leads to the appealing result that someone with a bequest motive will hold more wealth than someone without a bequest motive. The result is derived in the context of mortality, health, and interest rate uncertainty, but similar results could be obtained allowing for other kinds of uncertainty. I also present data that show the retired elderly in my sample do dissave. I conclude that, in con-

[1]Bernheim calculated the 1969–75 wealth changes over just 574 households, and the 1975–79 changes over 1047 households. From the same data set I use an average of 2071 households over the first period and an average of 3673 households over the second period.

tradiction to many previous studies, the wealth-age relationship of the elderly is consistent with the strict life cycle hypothesis. Then I test for a bequest motive. My test is whether the saving of the elderly who have living children differs from the the saving of the elderly who do not have living children. I find no evidence for a bequest motive.

I. A Test of a Bequest Motive

Someone is said to have a bequest motive if he (or she) cares about the welfare of the recipients of his estate. Whether he will desire to leave a bequest, however, depends on how, given his wealth, he compares the utility of his own consumption to the welfare of his heirs.[2] For example, even with a bequest motive, someone with little wealth may desire to consume all his wealth over his remaining lifetime. Desired bequests would be zero. Because the date of death is uncertain, however, most people will leave bequests, and, if they are very risk averse, the bequests could be large. Therefore, bequests, large or small, are not evidence either for a bequest motive or for desired bequests: to infer the importance of a bequest motive, one needs to show how a bequest motive affects observable variables, and then formulate a test based on those variables.

Intuition suggests that someone with a bequest motive will desire to hold more wealth than someone without a bequest motive, and therefore consume at a slower rate; in fact, one could define a bequest motive as a condition on the rate of consumption. It seems better, however, to define a bequest motive in terms of utility and then derive the implications for the consumption path. This is especially useful when there are random elements in the environment beyond mortality uncertainty. For example, one would want to show that even though there is substantial uncertainty about future health status, a bequest motive will still produce a definite prediction about the consumption path.

In the model of bequests I propose, expected lifetime utility depends on consumption at each future date. If someone has a bequest motive, his utility function has more terms because, should he die, his wealth will increase the well-being of someone about whom he cares. Additional determinents of expected lifetime utility are mortality risk, and other stochastic events which include health status and the rates of return on assets. I derive the result in the Appendix that, given initial wealth, someone with a bequest motive will desire to consume less than someone without a bequest motive when the distributions of the future random variables are the same for the type types of people; therefore, with the same resources, the wealth of someone with a bequest motive will decline more slowly than the wealth of someone without a bequest motive. The argument can easily be extended to compare someone with a weak bequest motive to someone with a strong bequest motive, and, therefore, to people with many levels of bequest motives.

It may be that everyone has at least a small bequest motive in that everyone cares about society. One could call this the minimum value of the bequest motive. It is not a reasonable empirical goal to estimate the effect on consumption of this minimum value because we do not know what a consumption path would be were there no bequest motive whatsoever. A reasonable empirical goal is to compare the consumption path of someone who cares strongly about the welfare of his heirs with the consumption path of someone who cares only weakly about the welfare of his heirs. A small difference in the consumption paths would be good evidence that, even though there may be substantial differences in the concern for the welfare of the heirs, people value their own utility from consumption highly enough that variations in the marginal utility of bequests do not change behavior; that is, people are at a corner solution in the tradeoff

[2] One might think that if people have no bequest motive they would fully annuitize their wealth. But in the *RHS* the amount of privately purchased annuities is very small even among people with no identifiable heirs. The most likely explanation is that the rates of return on annuities are very low, often dominated by long-term bonds (Benjamin Friedman and Mark Warshawsky, 1985).

between the utility their own consumption provides and the utility from a bequest. As in ordinary demand systems, the corner solution only implies an inequality in the marginal rate of substitution; one cannot learn its magnitude. An implication, however, is that if someone's wealth were to increase, actual bequests would not change.

Evidence for a bequest motive would be differences between the consumption paths of people who have differences in their concern for the welfare of their heirs. If there is no variation in consumption, one can only say there is no evidence for a bequest motive. However, one can conclude that a bequest motive is not an important determinant of consumption and wealth holdings, and that observed bequests are not the result of a desire to leave bequests; rather, they are caused by mortality uncertainty.

Of course, a difficulty with this method of testing for a bequest motive is that one must be able a priori to classify people according to the strength of their bequest motive. My original plan was to classify according to the number and type of heirs. Here I only classify according to whether the people have living children; the results did not warrant further investigation.

II. Empirical Findings

The first empirical result concerns the saving rate of the elderly: under the strict life cycle hypothesis (no bequest motive), one would expect the elderly to dissave as they age because of rising conditional mortality rates. The second result is a test of a bequest motive as specified in Section I. I study the wealth of the elderly as a group. Thus, I do not investigate individual behavior.[3] My results can best be compared with the results from cross-section analysis and from simulations.

A. *Data*

The data are from the *Longitudinal Retirement History Survey* (*RHS*).[4] About 11,000 households whose heads were born between 1906 and 1911 were interviewed every two years from 1969 through 1979. The *RHS* includes questions about all assets and liabilities with the exception of a meaningful question on the asset value of life insurance.[5] From the questions one can construct an (almost) complete balance sheet of the household.

The object of study is the change over a two-year period in bequeathable wealth. Data definitions are given in the Appendix; I mention here that the important components of bequeathable wealth are housing wealth, stocks and bonds, property, businesses and savings accounts less debts. I study bequeathable wealth rather than a broader wealth definition because typically the evolution of bequeathable wealth gives information about the utility function. Total expected lifetime resources, which are the sum of bequeathable wealth and the expected present value of annuities including Social Security and Medicare/Medicaid, do not give direct information about the utility function because they evolve according to a combination of individual choice and the mortality tables. The reasoning is that annuities are income which cannot be borrowed against. Therefore, the expected present value of the annuity stream does not enter the lifetime budget constraint as a stock; rather, the income flow acts as a boundary condition on the utility-maximization problem.[6] For example, someone whose only resource is Social Security is not free to consume more than the income from Social Security. The expected present value of Social Security will evolve according to the mortality tables; it

[3]King and Dicks-Mireaux, Kurz (1985), and Diamond and Hausman emphasize the heterogeneity of the elderly. There is certainly substantial variation in wealth holdings in this data set (see my paper with John Shoven, 1985).

[4]The *RHS* was commissioned by the Social Security Administration and conducted by the Census Bureau. See Lola Irelan (1972).

[5]My paper with Shoven describes the categories in detail.

[6]Further discussion of this point is in my 1986 paper.

TABLE 1—REAL WEALTH CHANGES FROM 1969 TO 1979

Initial Wealth	Population	Housing Wealth[a] Included	Housing Wealth[a] Not Included
Observations with Positive Initial Wealth	Singles	−25.2	−39.8
	Couples	−2.9	−16.9
	All	−15.0	−29.2
All Observations	Singles	−22.4	−36.4
	Couples	−2.0	−14.5
	All	−13.9	−27.3

Source: Author's calculations from the *RHS*.
[a]Shown in percent.

gives no information about the utility function.

To study wealth changes, one would estimate the coefficient in the equation $w_2 = kw_0$ in which w_2 and w_0 are real wealth levels in year 2 and year 0, respectively, and k is the wealth retention rate. In the *RHS* data, there appear to be observation errors in wealth, so I use an estimator that is consistent even when wealth is observed with error. I estimate k by

$$(1) \qquad \hat{k} = \sum w_2 / \sum w_0.$$

Neither the ratio estimator $\tilde{k} = 1/n\Sigma(w_2/w_0)$ nor OLS is consistent for k.

For singles and for couples I calculated k in each of the five two-year periods.[7] Each k was based on a different sample because of retirement and mortality. The ten-year wealth retention rate is the product of the k's: it gives the fraction of a dollar that would remain at the end of the ten years in real terms.

B. *Results*

Table 1 gives percentage changes in real wealth over the ten-year period of the *RHS*.[8] The columns show wealth changes according to whether or not housing wealth is included in bequeathable wealth. The first three rows are over observations which have positive bequeathable wealth in the initial period. The second three rows are over all observations. Table 2, to be discussed later, gives information on the number of observations behind the calculations.

In principle, all types of bequeathable assets will change as the consumption trajectory evolves: in practice, it is difficult to change the consumption level of housing because of the costs of transition from one consumption level to another. This is particularly true for the elderly. Until a complete model of desired housing services and transactions costs is developed, probably the best that can be done is to exclude housing wealth from the bequeathable wealth totals.[9] It is excluded in the wealth calculations that appear in later tables; but it turns out that no substantive conclusion is changed by including housing in bequeathable wealth.

The idea behind restricting the sample to include only observations with positive wealth is that households with little wealth will not follow desired wealth trajectories

[7]Were I to study wealth changes over periods longer than two years, the sample would be reduced due to mortality. Furthermore, the estimation should allow the households to reoptimize every two years in response to windfall gains and losses. I select only retired households: the theory is not easily testable if workers are included in the sample because one does not know their lifetime resources. See the Appendix for the selection criteria.

[8]The wealth levels in all tables were deflated by the CPI to find real wealth changes. Deflating by a cost-of-living index that is tailored to the elderly changes the results by very little. For example, the Boskin-Hurd index (Michael Boskin and myself, 1985) which is defined for five age groups of the elderly gives slightly less inflation than the CPI over the 10-year period (6.7 vs. 7.1 percent per year). This produces a rate of wealth change of −24.6 percent against −27.3 percent in Table 1 in the base case (no housing wealth, all observations).

[9]King and Dicks-Mireaux advocate a similar approach.

because they will have reached a borrowing constraint before two years have passed: once bequeathable wealth has been exhausted, consumption will be constrained to follow annuity income. The initial unconstrained rate of change of bequeathable wealth would be mismeasured when calculated over the entire two-year period. Furthermore, anyone with negative wealth is, in the context of the economic model, observed with error. Simple errors in variables arguments predict, however, that limiting the sample to positive initial wealth causes the rates of change to decrease which, indeed, is what is found in the first rows of Table 1. I believe that at this stage of descriptive statistics it is better to allow negative wealth than to predispose the wealth changes to be negative; thus, in later results I use the complete sample.

Table 1 shows that the elderly in this sample dissaved over the period 1969 to 1979: the estimates range from 13.9 to 29.2 percent of initial bequeathable wealth. In the case that I believe is most representative of desired wealth changes (housing wealth excluded, all observations), there is dissaving of 27.3 percent, which is at a rate of 3.2 percent per year. Both couples and singles dissave, singles more than couples. Couples are substantially different from singles in mortality rates (the household composed of a couple will survive longer, possibly not intact, than the household composed only of a single person), initial wealth, annuities, and in the frequency of children. Some of these explanations for the difference in rates of dissaving will be explored later.

Table 2 shows percentage changes in real bequeathable wealth, excluding housing, in each of the two-year periods, and the number of observations. Real wealth declined in all years except 1977–79. The table emphasizes an important fact: all the wealth changes in this paper are *ex post* wealth changes. The theory refers to desired or *ex ante* wealth changes. While one would expect the two to be equal on average, in any time period they will differ due to unanticipated windfall gains and losses. Apparently there were extraordinary losses in 1975–77 and extraordinary gains in 1977–79. In fact, the wealth changes in the two time periods average to about −7.2 percent per period

TABLE 2—REAL WEALTH CHANGES BY YEAR AND NUMBER OF OBSERVATIONS[a]

Year	Singles	Couples	All
1969–71	−3.9	−3.0	−3.6
	(1009)	(419)	(1428)
1971–73	−6.1	−2.5	−4.2
	(1290)	(740)	(2030)
1973–75	−12.6	−0.5	−7.3
	(1552)	(1204)	(2756)
1975–77	−19.7	−25.4	−22.3
	(1864)	(1511)	(3375)
1977–79	1.0	22.9	10.9
	(2187)	(1790)	(3977)

Source See Table 1.
[a]Shown in percent. The number of observations is shown in parentheses.

TABLE 3—REAL WEALTH CHANGES FROM 1969 TO 1979: BEQUEST MOTIVE[a]

	Living Children	No Living Children	All
Singles	−38.0	−32.6	−36.4
	(1104)	(477)	(1581)
Couples	−16.8	−1.7	−14.5
	(957)	(175)	(1132)
All	−28.2	−24.2	−27.3
	(2061)	(652)	(2713)

Source: See Table 1.
[a]Shown in percent. The average number of observations in each two-year period is shown in parentheses.

(geometrical average), which is a reasonable continuation of the rates in the three periods from 1969 to 1975. An investigation of the components of the losses and gains in the portfolios of the *RHS* households deserves attention, but it is beyond the scope of this paper.

Table 2 reveals a trend toward increasing rates of dissaving as the population ages and faces higher conditional mortality rates. Simple models of intertemporal utility maximization under mortality risk predict such a trend: an increase in the conditional mortality rate increases present consumption at the expense of future consumption (Menahem Yaari, 1965). However, these trends in wealth changes can only be suggestive because neither wealth nor annuities are constant over the time periods.

Table 3 gives wealth changes according to whether the household has living children or

not. Because of the age of the *RHS* sample, almost all of these children will be adults in their own households.[10] According to the theoretical results, a bequest motive will decrease consumption, and, therefore, the household will save more. The empirical result in the table is that households with children actually save less than households without children. This result does not depend on whether housing wealth is included: both singles and couples with children still save less than singles and couples without children.

Although the results in Table 3 give no evidence for a bequest motive, they are not conclusive because no account is taken of initial wealth and annuities. Over the five sample periods, households without children averaged substantially more bequeathable wealth than households with children; their annuities were about the same. In a simplified model, Bernheim shows that a high ratio of annuities to wealth causes a large percentage decline in wealth. This suggests that, even with a bequest motive, households with children might dissave more in percentage terms than households without children, and it certainly leaves open the possibility that if the wealth levels were the same, households with children would have higher savings rates.

To hold approximately constant the initial wealth and annuity levels, I divided the 1969 sample of couples with children into 16 cells according to the initial wealth quartile and annuity quartile, and calculated $\Sigma w_2/\Sigma w_0$ in each cell to give wealth retention rates by wealth and annuity quartile.[11] A similar calculation using the same quartile points was made over the 1969 sample of couples without children. The calculations were repeated for 1971, 1973, 1975, and 1977. Because the number of observations in some cells is small, and some initial wealth levels are close to zero, it is not meaningful to average the savings rates in each cell across years. Instead, in each of the cells I count the number of years in which the wealth retention rate of households with children exceeded the wealth retention rate of households without children. That count is recorded in Table 4. The greatest possible entry is 5. Under the hypothesis of no bequest motive, 2.5 is expected; under the alternative of a bequest motive, higher numbers are expected. The greatest possible entry in the final column or row is 20; with no bequest motive, 10 is expected.

A total of 160 comparisons were made. In about 47 percent of the comparisons (75/160), households who have children saved more than households who do not have children. This result is, of course, not expected under the hypothesis of a bequest motive. There is some difference by marital status; but, as shown in the next table, that is due to the method of classification.

Modigliani (1986) has suggested that desired bequests will increase sharply with wealth level. In Table 4, however, there seems to be no pattern over couples either by wealth level or by annuity level. Over singles, it appears there is some differential by wealth level: in the two lowest wealth quartiles, the cell comparisons favor a bequest motive in 26 of 40 cases, or about 65 percent. Singles in those wealth quartiles are poor: for example, in 1975 the quartile points for singles were \$1200, \$5759, and \$18000, excluding housing wealth. Singles with children would mostly be widows. Although the annuity variable includes transfers from relatives, it may be that there are more unrecorded transfers from children in the lowest quartiles than in the highest quartiles. The effect is not strong enough in the table to draw any firm conclusion without more investigation.

A problem with the classification method in Table 4 is that with observation errors on w_0, wealth retention rates are bound to be higher in the lower wealth quartiles than in the higher wealth quartiles. Put differently, the estimator given in (1) is not consistent

[10] No information is given on the children, but most are probably between 30 and 45-years-old. Thus, very few households have children at home. Excluding households in which children are present does not change the basic results.

[11] The annuity classification is only approximate. According to the basic theory the entire annuity trajectory influences the wealth path. I used annuity wealth to reduce the trajectory to a single number, which was used for the classification. This is preferable to classifying by annuity income in that some early retirees must wait several years to begin to receive Social Security and private pensions.

TABLE 4—COMPARISON OF SAVING RATES BY INITIAL WEALTH AND ANNUITY QUARTILES

Wealth Quartiles	Annuity Quartiles 1	2	3	4	All Annuity Levels
A. Couples					
1	2/5	2/5	4/5	2/5	10/20
2	1/5	1/5	3/5	2/5	7/20
3	1/5	4/5	2/5	1/5	8/20
4	2/5	2/5	2/5	2/5	8/20
All Wealth Levels	6/20	9/20	11/20	7/20	33/80
B. Singles					
1	3/5	3/5	2/5	3/5	11/20
2	5/5	3/5	4/5	3/5	15/20
3	2/5	2/5	0/5	3/5	7/20
4	3/5	2/5	1/5	3/5	9/20
All Wealth Levels	13/20	10/20	7/20	12/20	42/80

Source: See Table 1.
Note: Entries are the fraction of years in which saving rates of households with children exceeded the saving rates of households without children.

TABLE 5—COMPARISON OF SAVING RATES BY INITIAL CAPITAL INCOME AND ANNUITY QUARTILES

Income Quartiles	Annuity Quartiles 1	2	3	4	All Annuity Levels
A. Couples					
1	2/5	2/5	3/5	2/5	9/20
2	2/5	2/5	3/5	2/5	9/20
3	1/5	2/5	3/5	1/5	7/20
4	2/5	2/5	1/5	4/5	9/20
All Income Levels	7/20	8/20	10/20	9/20	34/80
B. Singles					
1	3/5	0/3	2/5	2/5	7/18
2	2/5	2/5	2/4	2/4	8/18
3	2/5	2/5	3/5	3/5	10/20
4	3/5	2/5	2/5	0/5	7/20
All Income Levels	10/20	6/18	9/19	7/19	32/76

Source and *Note*: See Table 4.

for w_2/w_0 in each cell when observations are assigned according to w_0. Furthermore, the means in the lowest quartile will be small making the variance of $\Sigma w_2/\Sigma w_0$ large. Table 5 gives similar results but the classification is by quartiles of initial capital income rather than by quartiles of initial capital wealth.[12] Again, the quartiles for couples are calculated across the capital income of all couples whether or not the household has living children, and similarly for singles.

[12] This way of classification is like instrumental variables classification: under instrumental variables, the classification would be from fitted values of the probability that a household fell in a particular cell where the predictor would be capital income. When w_0 has observation error, capital income is a good instrumental variable because it comes directly from the survey data; it is not derived from capital.

The results for couples are very similar to those given in Table 4. As before, there is no pattern by initial capital income or annuity wealth. Over singles, the differences by initial wealth that were found in Table 4 have disappeared, and there is no longer any evidence for speculation about unrecorded family transfers.[13] As with couples, it appears that any differential saving according to whether the household has children or not is random with respect to annuities and capital income. One difference from Table 4 is that the fraction of cells in which saving was higher for single households with children fell to 0.42. which is very close to the fraction for couples. The general impression, as in Table 3, is that there is no evidence for a bequest motive even when wealth and annuities are held constant.

III. Conclusions

Over the five two-year periods of the *RHS*, the elderly in the sample generally decumulated real wealth. The estimated rate of decumulation over ten years is about 3.2 percent per year. At this rate, a household with a twenty-year life expectancy will have reduced its bequeathable wealth to about half of its initial level. However, theory suggests that the rate of decumulation will increase with age, so one would expect the wealth level to be less than half after twenty years.

These results are in contradiction to most cross-section results, but I believe these results are more reliable. First, I study only the wealth changes of the retired elderly, so that no speculation about lifetime earnings is necessary. I can calculate total resources and, with time separability of the utility function, no further data on past earnings are required to find the desired wealth trajectory. Second, I study the wealth changes of individuals over time rather than having to compare the wealth levels of different people. Third, and I believe most importantly, differential mortality by wealth level causes cross-section wealth comparisons of the elderly to be almost meaningless. I offer the following evidence to support this statement. Between 1969 and 1979 the median wealth of all retired couples and of all retired singles grew substantially in real terms; yet, in every two-year period the median wealth of each group, holding composition constant, fell. The reconciliation of these two apparently contradictory facts is that in every period, the initial wealth of the couples who survived the period was substantially higher than the initial wealth of the couples who did not survive; the median on average was about 54 percent higher. Thus, the median wealth of all couples was higher at the end of the period than at the beginning precisely because couples with below-average wealth left the sample of couples. Furthermore, the survivors, mostly widows, had higher end-of-period wealth than the end-of-period wealth of the singles, about twice as much, so that the median wealth of the singles increased over the period. Therefore, differential mortality by wealth level caused the wealth of both groups to rise, even though the wealth of individual intact households fell. A comparison of median wealth in 1969 to median wealth in 1979 roughly duplicates what is done in cross section. Because of changes in composition, conclusions based on the medians of couples and of singles would be completely misleading about individual behavior.

There is no evidence for a bequest motive, at least insofar as it depends on whether the household has living children. This does not necessarily mean that parents do not care about the welfare of their children. In fact, the wealth data suggest that parents transfer substantial wealth to their children, but, as would be suggested by a standard human capital model, it is transferred earlier in life to support the children's consumption and education. For example, in 1975, couples with children had initial wealth excluding housing equity of about $32,000 whereas couples without children had initial wealth of about $47,000.[14] For singles, the corre-

[13] There were no single households in 4 cells; thus, there are only 76 comparisons.

[14] Of course, the comparison does not hold lifetime wealth constant.

sponding figures are \$10,000 and \$19,000. If housing is included, the figures for couples and singles are raised about \$20,000 and \$10,000, respectively. Furthermore, the children belong to a generation with high expected lifetime wealth. Apparently the transfers that have already been made and the bright prospects of the children cause parents to choose consumption rates that are indistinguishable from the rates of people of similar wealth levels but who have no children.

Kotlikoff and Summers estimate that about 80 percent of the capital stock held by households arises from intergenerational transfers. The results of this paper cannot be used to check that estimate because the wealth holdings of the *RHS* sample cannot be aggregated to estimate wealth holdings of the population. Nonetheless, these results do have implications for the Kotlikoff-Summers findings.

Even though no bequest motive was detected by the methods of this paper, bequests could be a superior good to such an extent that only the very wealthy respond to the bequest motive. In that the distribution of wealth is highly skewed, a few large desired bequests could account for most desired bequests, and be an important part of capital transfers.[15] Surveys like the *RHS* are unlikely to be useful to study desired bequests of the very wealthy simply because the extremely wealthy are probably reluctant to be interviewed.[16] It should be noted, however, that even in the upper wealth quartile there was no evidence for a bequest motive. One would imagine that even if only a few wealthy in the *RHS* had a bequest motive, it would be detected in the upper wealth quartile: the estimator of the wealth retention rate in each cell can be written as

$$(2) \qquad \hat{k} = \sum w_2 / \sum w_0 = \left(\sum (w_2/w_0) w_0\right) / \sum w_0 .$$

[15]In 1979, households in the upper 10 percent wealth tail held about 46 percent of bequeathable wealth. If one excludes housing, the figure rises to 55 percent (see my paper with Shoven).

[16]A further problem in the *RHS* is that the maximum entry in any asset category is \$999,999.

TABLE 6

	Living Children		No Living Children
	No Transfers	All[a]	All[a]
Couples	−13.5	−16.8	−1.7
	(769)	(957)	(175)
Singles	−41.1	−38.0	−32.6
	(782)	(1104)	(477)

Note: Shown in percent. The average number of observations is shown in parentheses.

[a]From Table 3.

This is a weighted average of individual rates where the weights are initial wealth. One would expect the estimate of k in the upper wealth quartile to be dominated by the wealth changes of the very wealthy.

Intervivos giving could be an important part of intergenerational transfers. The *RHS* has questions on amounts given to relatives and children outside the home. The amounts are very small, ranging from \$39 to \$60 on average, depending on the year. While these transfers are probably highly concentrated and may be important to a few individuals, they are too small to affect average rates of decumulation. The *RHS* also has questions on the number of children supported either fully or partially, and on whether support is received from children. To find whether intergenerational transfers might affect estimated saving rates, I estimated the wealth retention rates over the sample which neither supports children nor is supported by them. The ten-year rates of wealth change for that sample along with some excerpts from Table 3 for comparison are shown in Table 6. There is no change in the basic result: eliminating households in which there are transfers between the parents and the children increased the measured saving rate for couples and decreased it for singles, but the saving rates of households without children remain higher than the saving rates of households with children. The change in the number of observations indicates that there are substantial numbers of families that have some transfers, but apparently the magnitude of the transfers is small.

The most straightforward interpretation of the results of this paper is that in the *RHS*

any bequest motive is not an important determinent of consumption decisions and wealth holdings. Because the *RHS* is a self-weighting sample of heads of households, and it spans ages 58–74, the results probably extend to the entire elderly population with the possible exception of the very wealthy. Bequests seem to be simply the result of mortality risk combined with a very weak market for private annuities. If this is the case, there is no reason to replace the strict life cycle hypothesis by models that emphasize the determinents of intergenerational transfers, as called for by Kotlikoff and Summers. Of course, one should use a model that illuminates the question under study. If one is interested in understanding how most elderly would respond to, say, a change in Social Security benefits, the life cycle hypothesis, possibly augmented for health and interest rate uncertainty, is surely the place to start. If one wants to understand how the capital stock is accumulated, one would probably want to study the very wealthy. However, the standard consumption models may not apply: time constraints prevent the very wealthy from consuming even the interest from their wealth.

Appendix

A. *The Effect of a Bequest Motive on Wealth Holdings*

I assume there are two kinds of consumers: those with a bequest motive ($B=1$) and those without a bequest motive ($B=0$). Given initial wealth, w_t, each desires to maximize the expected value of lifetime utility in the face of uncertainty about the date of death, the rate of return on wealth, and health status. Although there are other ways to model the influence of health on consumption, I assume that it affects the utility of consumption.

I specify a finite time horizon divided into N periods. As in any period analysis, there is some arbitrariness about the sequence of events in a period. I assume that given w_t and health status, h_t, consumers choose consumption, c_t. Having consumed c_t, they die with probability m_t, and leave $w_t - c_t$ to any heirs or to society. If they do not die, they carry over to the next time period $w_{t+1} = (w_t - c_t)(1+r_t)$, where r_t is stochastic when the decision on c_t is made.

The consumer problem is to maximize in c_t

$$\text{(A1)}\quad U(c_t) + m_t V(w_t - c_t;\ B) + (1-m_t)E_t[\Omega_{t+1}(w_{t+1}, h_{t+1};\ B)],$$

where $U(\cdot)$ is the utility from consumption, $V(\cdot;\ \cdot)$ is the utility from bequests, Ω_{t+1} is the maximum expected lifetime utility given w_{t+1}, and h_{t+1} is the vector of random outcomes on future health. At all times I compare two consumers with the same realizations on h_t, so I have suppressed h_t as an argument in $U(\cdot)$. Expectations, $E_t[\cdot]$, are taken before observations on mortality, r_t, and h_{t+1} have been made. Ω_{t+1} incorporates future decisions about the allocation of wealth between consumption and saving, so it will depend on B. It can be defined recursively as

$$\text{(A2)}\quad \Omega_{t+1}(w_{t+1};\ B) = \max\{U(c_{t+1}) + m_{t+1}V(w_{t+1} - c_{t+1};\ B) + (1-m_{t+1}) \times E_{t+1}[\Omega_{t+2}(w_{t+2}), h_{t+1};\ B)]\},$$

where the maximizations is in c_{t+1}. I assume the utility functions have properties $U' > 0$, $U'' < 0$, and $U' \to \infty$ $c \to 0$; $V' = 0$ when $B=0$; $V' > 0$ when $B=1$; $\Omega'_{t+1} > 0$, $\Omega''_{t+1} < 0$. The notation U' means the partial derivative with respect to the first argument, and similarly for V' and Ω'; U'' and Ω'' mean the second derivatives. Let $c_t^*(B)$ be the solution to the consumer problem. $c_t^*(B)$ will depend on B, and it will be a function of w_t. I have supressed w_t and h_t since for this analysis they will be the same for both types of consumers. I want to show that $c_t^*(0) > c_t^*(1)$, and, therefore, that $w_{t+1}^*(1) = (w_t - c_t^*(1))(1 + r_t) > w_t^*(0) = (w_t - c_t^*(0))(1 + r_t)$ when the realizations on r_t are the same. If the distribution of r_t is the same for both types of consumers, $E_t[w_{t+1}^*(1)] > E_t(w_{t+1}^*(0))$. Then, the inequality will hold for the average

wealth of both types of consumers if the distribution of health is the same.

The method of proof will be to show that a bequest motive increases the marginal utility of wealth, Ω_t; that is, $\Omega_t'(w_t;\ 1) > \Omega_t'(w_t;\ 0)$ for $t < N$. This is certainly reasonable in that a bequest motive means wealth produces utility both when the wealth holder dies and when he lives. Given the result on the marginal utility of wealth, the conclusion will follow easily. The inequality on the marginal utility of wealth will be demonstrated by mathematical induction. Thus I first show that $\Omega_N'(w_N;\ 1) \geqq \Omega_N'(w_N;\ 0)$.

$$\text{(A3)}\quad \Omega_N = \max_{c_N} \{U(c_N) + V(w_N - c_N;\ B)\},$$

subject to $c_N \leqq w_N$. The first-order conditions for this problem are

$$\text{(A4)}\quad (U' - V') \geqq 0, \quad \text{and} \quad = 0 \text{ if } c_N^* < w_N.$$

If $B = 0$, $V' = 0$ and $c_N^* = w_N$. If $B = 1$, two solutions are possible: The first is that $c_N^* = w_N$ and $V'(0;1) < U'(w_N)$; that is, the marginal utility of bequests is so small that no wealth is allocated to bequests in the last period. One wants to allow a corner solution to be possible: the utility from bequests should not become unbounded as bequests become small because the heirs have their own resources.

The second solution obtains when w_N is large; then $c_N^* < w_N$ and $U'(c_N^*) = V'(w_N - c_N^*;\ 1)$. Then from the first-order conditions,

$$\begin{aligned}\text{(A5)}\quad & \Omega_N'(w_N;\ 1) \\ & = U'\,\partial c_N^*/\partial w_N + (1 - \partial c_N^*/\partial w_N)V' \\ & = (U' - V')\,\partial c_N^*/\partial w_N + V' \\ & = V' = U'(c_N^*).\end{aligned}$$

The second solution implies that

$$\begin{aligned}\text{(A6)}\quad & \Omega_N'(w_N;\ 0) = U'(w_N) \\ & < U'(c_N^*(1)) = \Omega_N'(w_N;\ 1)\end{aligned}$$

in that $c_N^*(1) < w_N = c_N^*(0)$.

This establishes, then, that in the last period the marginal utility of wealth of consumers with a bequest motive is at least as great as the marginal utility of wealth of consumers without a bequest motive.

The next step in the mathematical induction is to show that if

$$\text{(A7)}\quad \Omega_{t+1}'(w_{t+1};\ 1) \geqq \Omega_{t+1}'(w_{t+1};\ 0),$$

then $\qquad \Omega_t'(w_t;\ 1) > \Omega_t'(w_t;\ 0).$

The proof is as follows:

$$\begin{aligned}\text{(A8)}\quad \Omega_t = \max_{c_t} \{ & U(c_t) + m_t V(w_t - c_t;\ B) \\ & + (1 - m_t) E_t[\Omega_{t+1}(w_{t+1}, h_{t+1};\ B)]\}.\end{aligned}$$

The first-order conditions include

$$\begin{aligned}\text{(A9)}\quad & U' - m_t V' - (1 - m_t) \\ & \times E_t[\Omega_{t+1}'(w_{t+1}, h_{t+1};\ B)(1 + r_t)] = 0,\end{aligned}$$

assuming an interior solution, $0 < c_t < w_t$. Let $c_t^*(0)$ be the solution to this problem for $B = 0$, and $c_t^*(1)$ be the solution for $B = 1$. That is,

$$\begin{aligned}\text{(A10)}\quad & U'(c_t^*(0)) = (1 - m_t) \\ & \times E_t[\Omega_{t+1}'(w_{t+1}^*(0), h_{t+1};\ 0)(1 + r_t)],\end{aligned}$$

where $w_{t+1}^*(B) = (w_t - c_t^{*\prime}(B))(1 + r_t)$. If $c_t^*(1) < c_t^*(0)$ it will be easy to get the desired result. Suppose, however that $c_t^*(1) \geqq c_t^*(0)$. Then

$$\begin{aligned}\text{(A11)}\quad & U'(c_t^*(1)) \leqq U'(c_t^*(0)) \\ & < (1 - m_t) E_t[\Omega_{t+1}'(w_{t+1}^*(0), h_{t+1};\ 0) \\ & \quad \times (1 + r_t)] + m_t V'(w_t - c_t^*(1);\ 1) \\ & \leqq (1 - m_t) E_t[\Omega_{t+1}'(w_{t+1}^*(0), h_{t+1};\ 1) \\ & \quad \times (1 + r_t)] + m_t V'(w_t - c_t^*(1);\ 1) \\ & < (1 - m_t) E_t[\Omega_{t+1}'(w_{t+1}^*(1), h_{t+1};\ 1) \\ & \quad \times (1 + r_t)] + m_t V'(w_t - c_t^*(1);\ 1)\end{aligned}$$

But the first-order conditions for type 1 consumers require that the left-hand side of (A11) equals the right-hand side of (A11). Therefore, it must be that $c_t^*(1) < c_t^*(0)$.

Then,

$$\text{(A12)} \quad \Omega_t'(w_t;\, B) = U'\partial c_t^*/\partial w_t + m_t(1-\partial c_t^*/\partial w_t)V'(w_t - c_t^*(B);\, B) + (1-m_t)(1-\partial c_t^*/\partial w_t) \times E_t\left[\Omega_{t+1}'(w_{t+1}^*, h_{t+1};\, B)(1+r_t)\right].$$

From the first-order conditions

$$\begin{aligned}\text{(A13)} \quad \Omega_t'(w_t;\, 0) &= (1-m_t) \times E_t\left[\Omega_{t+1}'(w_{t+1}^*, h_{t+1};\, 0)(1+r_t)\right] \\ &= U'(c_t^*(0)) < U'(c_t^*(1)) \\ &= m_t V'(w_t - c_t^*(1);\, 1) + (1-m_t) \times E_t\left[\Omega_{t+1}'(w_{t+1}^*, h_{t+1};\, 1)(1+r_t)\right] \\ &= \Omega_t'(w_t;\, 1).\end{aligned}$$

Thus, by mathematical induction, $\Omega_t'(w_t;\, 1) > \Omega_t'(w_t;\, 0)$ for all $t < N$.

The proof showed that if $\Omega_{t+1}'(w_{t+1};\, 1) > \Omega_{t+1}'(w_{t+1};\, 0)$ then $c_t^*(1) < c_t^*(0)$, always assuming that w_t and h_t are the same for both types of consumers. Therefore,

$$\begin{aligned}\text{(A14)} \quad w_{t+1}^*(1) &= (w_t - c_t^*(1))(1+r_t) \\ &> (w_t - c_t^*(0))(1+r_t) \\ &= w_{t+1}^*(0)\end{aligned}$$

for the same realizations on r_t. Finally,

$$\text{(A15)} \quad E(w_{t+1}^*(1)) > E(w_{t+1}^*(0)),$$

when the distribution of r_t is the same for both types of consumers. If the distribution of health is the same for both types of consumers, this inequality will hold for the expected wealth of each type. Thus, $w_{t+1} - w_t$ will, on average, be larger for someone with a bequest motive than for someone without a bequest motive.

B. *Description of the Data*

1) *Retirement History Survey*. The *RHS* is a self-weighting sample of noninstitutionalized heads of households who were born in 1906–11 and survived until 1969. The heads or their surviving spouses were interviewed every two years through 1979. The initial sample was 11,152; the final sample was 7,325. The quality of the data seems to be as good as the CPS data: the sample was chosen from CPS rotation groups, and the interviews were conducted by the census. The data have been used extensively and successfully by many investigators.

2) *Bequeathable wealth*. The *RHS* includes very detailed questions on assets and liabilities. Respondents were asked either about actual holdings or to estimate market values. Fifteen responses were aggregated to form the following assets: net business wealth, net real property, net vehicle value, U.S. Savings Bonds, stocks and bonds, loans owned, checking and savings accounts. Debts were medical, store, bank, and debts to private individuals. Bequeathable wealth is the sum of these assets less the sum of the liabilities. In Table 1, net housing equity was added.

3) *Annuity wealth* is the sum of the expected present value of Social Security, as calculated from the Social Security law, data on Social Security earnings and mortality tables, Railroad Retirement, military, government, and private pensions, and the expected present value of transfers from relatives, Supplemental Security Income, welfare, Medicare and Medicaid, and private annuities. Discounting was 3 percent for real flows, and at a rate that depended on the corporate bond rate for nominal flows.

4) *Capital income* is the sum of interest and dividends, a service flow from housing equity, and rental income.

5) *Imputation methods*. Because there are more than 40 asset, liability, and income catagories, there are missing values. To eliminate observations on the basis of any missing values would be to reduce substantially the working sample; therefore, missing values were imputed. The imputation method is described in detail in my paper with John Shoven; but here I give a brief descrip-

tion. The goal of the imputation method was to retain information about the asset holdings of the individual. If a respondent indicated he had an asset but the amount was missing, other survey years were searched to find a valid value of the asset. A median rate of growth was applied to the valid entry of the individual to impute a value in the year in which it was missing. If no valid values could be found, the median over observations with positive values by marital status was imputed. If, in a particular year, a question about a particular asset was not asked, an interpolation for that individual from adjacent years was used. This did not happen for the important asset catagories.

6) *Sample selection.* A household was included in the sample for a two-year period if it was intact over the period, and if both the head of the household and his spouse had no earnings in that two-year period and no earnings any of the succeeding years of the survey.

REFERENCES

Bernheim, B. Douglas, "Dissaving After Retirement: Testing the Pure Life Cycle Hypothesis," in Zvi Bodi et al., eds., *Issues in Pension Economics*, NBER, Chicago: University of Chicago Press, 1987.

Boskin, Michael and Hurd, Michael, "Indexing Social Security Benefits: A Separate Price Index for the Elderly?," *Public Finance Quarterly*, October 1985, *13*, 436–49.

Danziger, Sheldon et al., "The Life-Cycle Hypothesis and the Consumption Behavior of the Elderly," *Journal of Post Keynesian Economics*, Winter 1982, *5*, 208–27.

Darby, Michael, *The Effects of Social Security on Income and the Capital Stock*, Washington: American Enterprise Institute, 1979.

Diamond, Peter and Hausman, Jerry, "Individual Retirement and Savings Behavior," *Journal of Public Economics*, February-March, 1984, *23*, 81–114.

Flavin, Marjorie, "The Adjustment of Consumption to Changing Expectations about Future Income," *Journal of Political Economy*, October 1981, *89*, 974–1009.

Friedman, Benjamin and Warshawsky, Mark, "Annuity Yields and Saving Behavior in the United States," presented at the NBER Conference on Pensions in the U.S. Economy, Baltimore, March 1985.

Hall, Robert E., "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, December 1978, *86*, 971–87.

______, "Real Interest and Consumption," NBER Working Paper 1694, 1985.

______, **and Mishkin, Frederic,** "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, March 1982, *50*, 461–81.

Hayashi, Fumio, "The Permanent Income Hypothesis: Estimation and Testing by Instrumental Variables," *Journal of Political Economy*, October 1982, *90*, 895–918.

______, "Tests for Liquidity Constraints: A Critical Survey," NBER Working Paper 1720, 1985.

Hurd, Michael, "Savings and Bequests," NBER Working Paper 1826, 1986.

______ **and Shoven, John,** "Inflation Vulnerability, Income, and Wealth of the Elderly, 1969–1979," in Martin David and Tim Smeeding, eds., *Horizontal Equity, Uncertainty, and Economic Well-Being*, NBER, Chicago: University of Chicago Press, 1985.

Irelan, Lola M., "Retirement History Survey: Introduction," *Social Security Bulletin*, November 1972, *35*, 3–8.

King, Mervyn, "The Economics of Saving: A Survey of Recent Contributions," in Kenneth Arrow and S. Hankapohja, eds., *Frontiers of Economics*, Oxford: Basil Blackwell, 1985.

______ **and Dicks-Mireaux, Louis,** "Asset Holdings and the Life-Cycle," *Economic Journal*, June 1982, *92*, 247–67.

Kotlikoff, Laurence, and Summers, Lawrence, "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 706–32.

Kurz, Mordecai, "Capital Accumulation and the Characteristics of Private Intergenerational Transfers," *Economica*, February 1984, *51*, 1–22.

______, "Heterogeneity in Savings Behavior:

A Comment," in Kenneth Arrow and S. Hankapohja, eds., *Frontiers of Economics*, Oxford: Basil Blackwell, 1985.

Menchik, Paul and David, Martin, "Income Distribution, Lifetime Savings and Bequests," *American Economic Review*, September 1983, *73*, 672–90.

Mirer, Thad, "The Wealth-Age Relation Among the Aged," *American Economic Review*, June 1979, *69*, 435–43.

Modigliani, Franco, "Life Cycle, Individual Thrift, and the Wealth of Nations," *American Economic Review*, June 1986, *76*, 297–313.

______ **and Brumberg, Richard,** "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in K. Kurihara, ed., *Post-Keynesian Economics*, New Brunswick: Rutgers University Press, 1954.

White, Betsey, "Empirical Tests of the Life Cycle Hypothesis," *American Economic Review*, September 1978, *68*, 547–60.

______, "Empirical Tests of the Life Cycle Hypothesis: Reply," *American Economic Review*, March 1984, *74*, 258–59.

Yaari, Menahem E., "Uncertain Lifetime, Life Insurance and the Theory of the Consumer," *Review of Economic Studies*, March 1965, *32*, 137–50.

Emigration to South Africa's Mines

By ROBERT E. B. LUCAS*

Temporary labor migration from five countries to South Africa's mines is examined. Emigration (a) *diminishes domestic crop production in the short run;* (b) *enhances crop productivity and cattle accumulation through invested remittances in the long run;* (c) *increases domestic plantation wages. Conflicting interests thus exist between employers in the sending countries and in the mines. State intervention adopted in the sending countries includes forced labor, emigration quotas, and compulsory population relocation.*

For more than a century the circular migration of younger men to the South African mines has profoundly affected the economies, political relations, and social structures of the countries of southern Africa. Indeed, the changes wrought by the continuing temporary emigration have led many observers to question the net benefits to the major labor supplying regions: Botswana, Lesotho, Malawi, Mozambique, and the South African "homelands."[1] The magnitude of miner recruitment and hence the potential effects on the labor supplying economies are certainly comparatively large. At about the time of the 1970 censuses, the stock of men employed in the South African mines relative to the *de jure* male population ages 18 to 35 exceeded 80 percent for Lesotho, was close to 50 percent for Botswana, and about 15 percent even for Mozambique where the proportion was lowest.

This paper presents a simultaneous, econometric model of both the determinants of international migration to the South African mines and of some of the economic consequences for each of the labor supplying countries.[2] Not only are the short-run effects of labor withdrawal on traditional crop cultivation and the domestic wage labor markets considered, but also the long-term effects of savings from mine earnings invested in crops and cattle in the home countries.[3]

A stylized model is outlined in Section I. However, variations in institutional arrangements, market conditions, and data avail-

*Economics Department, Boston University, 270 Bay State Rd., Boston, MA 02215. Initial funding for this study was provided by the ILO, Geneva. Much of the analysis and drafting were completed while I visited the Migration and Development Program at Harvard University. The African Studies Center at Boston University provided a travel grant to gather data in several countries in southern Africa. To the ILO, the Migration and Development Program, and to the African Studies Center I am extremely grateful for their generous support. All results and interpretations are solely my responsibility. My special thanks go to Oded Stark and two anonymous referees for detailed comments on an earlier draft.

[1]Some of these doubts are beyond the scope of the present study: the substantial external costs imposed by miners returning with tuberculosis or venereal diseases; the proliferation of single-parent households; the risks inherent in continued dependence upon South Africa; or the political costs to frontline nations in implicitly supporting the apartheid regime. However, some of the doubts also hinge on the more limited range of economic issues considered here. See, for example, Fion de Vletter (1981), Walter Elkan (1980), Elizabeth Gordon (1981), J. Halpern (1965), Donald Kowett (1978), Colin Murray (1981), Isaac Schapera (1947), Charles Stahl and Roger Bohning (1981), and Francis Wilson (1976).

[2]Although numerous internal migration functions have been estimated in various contexts, surprisingly little econometric evidence exists on the determinants of international migration for any region. Derek Byerlee (1972), Michael Greenwood (1975), T. Paul Schultz (1982), and Michael Todaro (1976) survey portions of the estimated internal migration equation literature. George Psacharopoulos (1976) and I (1976) present estimates of international migration parameters.

[3]Such long-run effects of out-migration have generally been neglected in the migration literature. See, however, Paul Collier and Deepak Lal (1980), myself and Oded Stark (1985), Henry Rempel and Richard Lobdell (1978), and Stark (1978, 1980).

ability require precise specification of the stylized model to differ from country to country, and these differences are discussed in Section II. Indeed, some of the policies adopted with respect to mine labor and associated markets have been so dramatic as to render the political economy of this market especially intriguing—the cutoff of all mine recruiting from Malawi in 1974, the use of forced labor in Mozambique under the Portuguese, or the forced exodus of South Africans to designated homelands.

The model is estimated over annual time-series data covering 1946 through 1978, the results being presented in Section III.[4] Some implications of these findings and closing remarks are contained in Section IV.

I. A Stylized Approach

The stylized model comprises four main blocks. The first deals with the determinants of migration to the South African mines from each country. In broad terms, the supply of miners is conceived to depend upon employment and earning opportunities at home, either in wage labor or family farming, relative to wages available at the mines. However, the actual flow of migrants may not be supply determined but rather dictated by policy-imposed limits in certain periods. The second block then models own-account crop production as affected both in the short run by labor withdrawal for mine and wage employment, and in the long run by investments saved or remitted by migrants. Third, investments in cattle herds, a major form of wealth in southern Africa, are modeled as a function of savings out of earnings. Lastly, the process of wage formation and extent of wage employment in the domestic labor markets are considered.

[4]I am particularly grateful to the following institutions for their hospitality and help while locating data: Botswana—National Archives, Central Statistics Office, Ministry of Agriculture, The University Library; Lesotho-National Archives, Bureau of Statistics; Malawi —National Archive Library, National Statistical Office, The University Library at Zomba; South Africa—Department of Statistics, Meteorological Office.

In a labor surplus economy, the blocks of this model would not be simultaneous: out-migration would affect neither agricultural production nor domestic wages and employment. But the stylized model allows for the possible absence of surplus labor and hence for simultaneity. Thus, out-migration might raise the domestic wage, diminish crop production in the short run but enhance it in the long run, and through each of these in turn affect the relative attraction of mine migration.

A. *Mine Migration*

In my article (1985a), I showed that, from 1974 onwards, the South African mining houses, acting in concert through their centralized recruiting agency, exercised a preference for domestic workers, where "domestic" refers to South Africa including the so-called Black States. Thus, after 1974, the market for foreign recruits did not operate on its supply curve. Moreover, as will be described in Section II, some of the sending countries imposed limits on the numbers of recruits during certain intervals. From any country, the number of men employed in the mines, m, is therefore determined through one of two processes: by equality with a quota, M, imposed by the sending country or, after 1974, by the mining houses; by a clearing process in a monopsonistic market with employment equal to supply.

In contrast to the model of John Harris and Michael Todaro (1970), no gamble with respect to likelihood or terms of employment prevails at the mine destination (though life, limb, and pride are certainly at risk). Recruits contract for a definite wage and fixed period of employment, after which they must be repatriated according to South African law.[5] The wage is known before migrating since recruitment occurs at depots scattered throughout the sending countries. Even working conditions are probably well understood by novices, for most are preceded by

[5]Until the mid-1970's, miners could not legally "desert" from a fixed-period contract. See Merle Lipton (1980).

relatives. On the other hand, to remain at home does involve a gamble, both in the chances of finding paid employment and of crop failure on one's own land. Reversing Harris-Todaro, but retaining their risk-neutrality assumption, the supply of miners from a given country may then be expressed as

$$(1) \qquad m = M \qquad \text{if a quota prevails,}$$

$$m(v_m - E\{y\}, n) \quad \text{otherwise,}$$

where $E\{y\}$ is the expected real income if one remains at home, v_m is the real mine wage, and n is male population within potential recruitment age.

Although some of the mine earnings are spent while in South Africa, a substantial fraction (approximately 60 percent on average) is remitted or sent as deferred pay for spending in the home country. The real mine wage is therefore measured by

$$(2) \qquad v_m = w_m/(.6p_h f + .4p_m),$$

where p_h and p_m are cost-of-living indices in the home country and in South Africa, respectively, w_m is the nominal mine wage and f a foreign exchange rate index between the South African Rand and local currency (where applicable).

Expected home income is derived from two main sources: from potential domestic wage employment and from returns to own-account agriculture.[6] The latter may be expressed as some proportion π of the value of the average product of labor in family farming, pq/l, where p is price of traditional crops, q is output, and l is labor involved in peasant agriculture. In addition, the expected wage if one seeks paid employment at home may be written $w_h e/(n-m)$ where w_h is the domestic nominal wage rate, and e is the extent of domestic male employment. Thus, the expected real income for those who remain in the country (expressed in local currency) is

$$(3) \quad E\{y\} = (w_h/p_h)(e/(n-m)) + \frac{p}{p_h}\pi\frac{q}{l}\phi\left[1 - \frac{e}{n-m}\right],$$

where l is a proportion ϕ of the nonemployed labor force, $n-m-e$. A linear specification of (1) may then be written for nonquota periods as

$$(4) \quad m = \mu_0 + \mu_1\left[v_m - \frac{w_h}{p_h\eta}\frac{e}{n-m}\right] + \mu_2\frac{p}{p_h}\frac{q}{n-m} + \mu_3 n$$

in which $\mu_2 = -\mu_1\pi/\eta$ and η is the base-year exchange rate. It might be expected that $\mu_1 \geq 0$, but a backward-bending supply response cannot be ruled out, a priori, particularly if miners are target savers (see Merle Lipton, 1980). Equation (4) gives the basic form of the migration function to be estimated here, but some modifications for particular circumstances of individual countries are necessary in Section II. In addition a semilogarithmic form is estimated, with $ln(m)$ as dependent variable, but also replacing $\mu_3 n$ with $\mu_3 ln(n)$ so that if $\mu_3 = 1$ this form of (4) could be rewritten with proportion of population migrating as dependent variable.

B. *Own-Account Crop Production*

Almost no prior estimates of crop production functions exist for this region and in particular for the tribal areas—the so-called Customary Land of Malawi, the "homelands" of South Africa, Tribal Lands of Botswana, and so forth (see, however, my paper, 1985c).

[6]No time-series data exist on nonagricultural self-employment of men. However, at least in parts of this region, much of the beer brewing and other forms of nonfarm self-employment are undertaken by women, though this may itself be a consequence of the long history of male absence at the mines. The six- and nine-month mine contracts traditional in Botswana, Lesotho, and Swaziland are designed to enable men to plough before signing on for the mines. Nonetheless, mine and crop work are conflicting alternatives, if only because arrival date of the rains is quite uncertain in many areas and it is essential to plow immediately once the rainy season commences. Many men sign on only to find ample rains arriving thereafter.

The chief difficulty is that, as in almost all developing countries, no capital stock series exists for this sector. Suppose, however, that some unknown fraction, σ_w, of wage earnings are saved and invested in equipment, working capital or technological improvement in traditional agriculture. Thus capital stock, k, in any period would be

$$(5)\quad k = k_{-1}[1-d] + \sigma_w\left[\frac{w_h}{p_h}e + v_m m\right],$$

where subscript -1 indicates a one-period lag and d is a rate of depreciation assumed equal to 0.1. From (5) a capital stock series denoted $kw = k/\sigma_w$ may be computed. However, this is not the only plausible source of savings for investment in crops: crop income itself obviously may also be capitalized in a similar fashion leading to a series denoted kq.[7] Given these capital stock surrogates, kw and kq, a crop output function of the following general form may then be explored in Section III:

$$(6)\quad lnq/N = \gamma_0 + \gamma_1 r + \gamma_2 ln(e+m)/N + \gamma_3 lnkw/N + \gamma_4 lnkq/N,$$

where N is total population, r is annual rainfall, and other variables are as before. Note that (6) has the advantage of emphasizing the short-run role of labor withdrawal through migration to the mines or internally for wage employment, and the long-run possibility of enhanced productivity through accumulation out of migrants' earnings. Clearly, the migration and production functions, (4) and (6), are simultaneously determined.

C. *Investments in Cattle*

In much of southern Africa, cattle are a major component in the traditional portfolio of assets. Several forms of investment are possible to accelerate the rate of growth in the herd: expenditures on improved veterinary techniques, selective breeding, import of prime beasts, installation of boreholes and fencing, or at least transitionally by postponing slaughter. To explore the contribution of migrants' savings to growth in the herd through such mechanisms, an approach related to (5) is adopted. In particular, let cw and cq be capitalized streams of wage and crop incomes, comparable to kw and kq but with a 20 percent rate of depreciation to allow for the shorter life span of cattle.[8] The desired stock of cattle, c^*, may be written as

$$(7)\quad c^* = \beta_0^* + \beta_1^* cw + \beta_2^* cq.$$

But ability to adjust quickly to a larger desired herd size is constrained by at least two factors:[9] the prior stock, both through calfing rate and survival (assuming some limitation on numbers imported); and variations in rainfall. Imposing a flexible accelerator modified by an additive rainfall effect, the observed number of cattle may be written

$$(8)\quad c = \beta_0 + \beta_1 cw + \beta_2 cq + \beta_3 r + \beta_4 c_{-1}.$$

D. *Domestic Wage Labor*

An a priori judgment must be made in each case whether the domestic wage labor market for men clears or even operates on its

[7] Initial values for the capital stock series are set by means of the formula: $k_0 = i_0/(d+g)$, where i_0 is investment in the initial period, g is the exponential growth rate of the investment series and d is the depreciation rate as before. This formula represents the stock which would have been reached given steady growth at rate g over an infinite horizon with depreciation as in (5). The growth rates are estimated by a simple regression on time over the first 10 years of the sample. In the case of wage earnings, i_0 is set equal to wage earnings in the first period of the sample. For crop income i_0 is measured by the average value of crops in the first 5 years since considerable fluctuations occur in some instances.

[8] Estimates of both the crop output and cattle herd equations with alternative depreciation rates adopted in constructing the various capitalized series (together with appropriately adjusted initial capital stocks) are also explored in Section III.

[9] Indeed, at least in Botswana, limited capacity in organized sector slaughtering has constrained even downward adjustments in herd size at certain peak times.

demand curve. In countries where much of the organized sector employment is in the public domain and parastatals are encouraged to provide jobs, both wage and employment are treated as exogenous. However, in the plantation or estate economies this is not true. Local variations are, however, paramount in formulating any approach to modeling an endogenous wage or level of employment and best considered within the context of specific countries in Section II.

II. The Specific Countries

In view of the diverse institutional arrangements and market conditions in each of the countries to be analyzed, and because of the variegated nature of each nation's data, the precise specifications of the stylized forms must differ from context to context. Unfortunately, it is quite impossible to itemize the extensive set of documents from which consistent time-series data were compiled for the present study, given the confines of space.[10] However, at least certain essential remarks on the data are made in developing the background to particular specifications.

A. *Malawi*

In April 1974, the government of Dr. Hastings Banda suspended further recruitment of miners. Ostensibly the reason was a plane crash, killing 74 returning Malawian miners. Though the 1974 cutoff was extreme, a precedent in milder form had existed during the colonial period. Prior to independence in 1964, the Nyasaland authorities imposed a quota on the number of men the mines' agents were entitled to recruit each year. The quota was varied from year to year, apparently depending upon the state of the domestic labor market, and was not always binding. By comparing the announced quota with recorded new recruits, years in which the quota actually restricted migration are distinguished. In accordance with (1), estimation of the mine labor supply equation is then confined to years prior to 1974 in which the quota was not binding.[11]

The distance from Malawi to South Africa is greater than from any of the other major suppliers of miners. Given the greater fixed cost of transporting the Malawi miners, the Employment Bureau of Africa (TEBA, the South African mines' centralized recruiting agency) has used a two-year contract for Malawi in contrast to a six- or nine-month term in Botswana, Lesotho, and Swaziland. The imposition of a two-year contract, combined with a Master and Servants law forbidding deserting that contract, has presumably dampened speed of adjustment to changing conditions for Malawians, and the number of mine workers lagged one year is therefore added to the migration equation (4).

Crop production on the Customary Lands of Malawi includes not only subsistence crops but also tobacco (see Edwin Dean, 1966). Productivity per acre in tobacco growing almost trebled from independence in 1964 to 1977 after far slower growth in the colonial period. In part, this may reflect the expanded use of fertilizers by smallholder farmers, so to the crop production function (6) is added a term for quantity of fertilizer sold to smallholders.[12] In addition, the number of tobacco growers on the Customary Lands has apparently been restricted, though whether this has been to exploit monopoly power in world trade in pipe tobacco or to protect estate producers is unclear. The effect on overall crop productivity of the permitted number of tobacco growers is explored by also adding this number to the crop output equation (6). But the potential restriction on tobacco growing may also affect mine labor migration which is no longer subject to free choice with respect to returns in crop cultivation as in (4). On the other hand, growing

[10] A detailed appendix is available from the author upon request, listing the data and itemizing the sources and precise definition of all variables in the model.

[11] The years thus excluded are 1951–52, 1954–56, 1958–59, and 1965. In these years the number of recruits actually slightly exceeded the announced quota by approximately 6 percent on average.

[12] A linear form, rather than logarithmic, is adopted since in earlier years fertilizer sales were zero.

of the principal subsistence crop, maize, has not been restricted, so for Malawi the crop value measure actually included in the mine labor supply equation is sales of maize, with tobacco represented by the potentially restricted number of growers.[13]

On the estates, the two major crops are tobacco and tea. The various estates compete amongst themselves for labor, but also compete with own-account agriculture, nonagricultural domestic employment, and the South African mines.[14] To examine the influence of these alternatives on estate work, and in particular to look at the effect of mine migration, structural equations for estate labor demand and supply are estimated. The specifications adopted respectively are

$$(9)\quad ln(e_a) = \delta_0 + \delta_1 ln\left(\frac{w_a}{p_a}\right) + \delta_2 ln(a_{te}) + \delta_3 ln(a_{to}) + \delta_4 ln(e_a)_{-1}$$

$$ln(e_a) = \lambda_0 + \lambda_1 ln\left(\frac{w_a}{p_h}\right) + \lambda_2 ln\left(\frac{w_{na}}{p_h}\right) + \lambda_3 ln(m) + \lambda_4 ln\left(\frac{p_c}{p_h}\right) + \lambda_5 ln(n_{to}) + \lambda_6 ln(N),$$

where w_a, e_a are the wage rate and number of employees in estate agriculture; p_a is a price index for tea and tobacco; a_{te} and a_{to} are the areas of estates under tea and tobacco, respectively; w_{na} is the wage rate in nonagricultural work; p_c is the price of maize; n_{to} is the number of tobacco growers, while N, m, and p_h represent population, miners, and home cost of living as before. The areas under tea and tobacco are included separately in the demand equation to allow for the difference in labor intensities between the two crops. Two of the alternatives to estate work—nonagricultural jobs and own-account maize cultivation—are represented by price terms in the supply equation: mine labor and tobacco growing on Customary Lands are not. Following the 1974 cutoff and during times of binding quotas in the colonial period, the mine wage was not a freely available alternative, and the number of mine workers is therefore included instead. Similarly, the suspected quantitative restrictions on the number of tobacco growers have already been mentioned. Population is included as a factor likely to have enhanced the supply of labor. Plausible hypotheses to be tested in Section III with respect to (9) are thus:

$$(10)\quad \delta_2, \delta_3, \delta_4, \lambda_1, \lambda_6 > 0$$
$$\delta_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 < 0.$$

In addition, tests with respect to flexible accelerator terms will also be reported in Section III.

B. *Mozambique*

Since Mozambique gained independence from the Portuguese in 1975, the South African mines have deliberately avoided contracting Mozambican novices, fearing political disruption in the mine compounds by workers from a Marxist state.[15] But disruption of mine recruiting preceded actual independence by several years as the armed conflict between FRELIMO and the Portuguese grew in severity.

In the 1970's, it thus seems migration from Mozambique to the South African mines has

[13] Data on sales of maize from Customary Lands are adopted as a proxy for value of production in the mine labor equation since output measures for maize are not available.

[14] Labor from Nyasaland also migrated to Northern Rhodesia and Southern Rhodesia prior to independence of the former as Zambia in 1964 and the unilateral declaration of independence in Rhodesia. However, the numbers involved were comparatively small and curtailed after the mid-1960's (Robert Boeder, 1974).

[15] Indeed, immediately following independence, the number of Mozambican miners dropped by half though it is unclear to what extent this reflected mistrust by the South Africans as opposed to initial disorganization in Mozambique and specifically the failure to issue appropriate passports (see W. J. Breytenbach, 1979, and Ruth First, 1983).

been shaped by quite different factors from those outlined in the stylized migration equation (4). A form of this equation is, however, explored for the earlier interval from 1946 through 1970, with four modifications.

(a) According to a 1942 circular issued by the Portuguese authorities, all working-age Africans were required to prove that they had worked for six months in every year (LeRoy Vail and Landeg White, 1980). The required employment could be satisfied voluntarily either through domestic wage work or by contracting with an authorized recruiter for work abroad, of whom the predecessor to TEBA, recruiting for the South African mines, was most important. Forced labor, at least for "vagrants," continued to be the subject of protest and investigation in the International Labour Organization throughout the 1950's (ILO, 1962; James Duffy, 1967). In 1961, a series of decrees altered the system of compulsory employment, and forced labor seems to have been significantly reduced thereafter, though even into the early 1970's reports of some forced labor continue. During the main period of compulsory registration and employment, the need to volunteer for work or face forced labor, under even worse wage and working conditions, is likely to have increased the propensity to migrate to the mines (Jeanne Penvenne, 1979). To test for this, a dummy variable is introduced into the migration equation (4), set equal to one for the period of extensive forced labor and registration up through 1961.

(b) The measures of employment included in defining the expected wage differential between Mozambique and the Rand mines are employment in Mozambican industry, mines and plantations. In addition, the construction sector has been a major source of employment, though no data on the number of workers involved could be located. As a proxy, construction output is therefore inserted into the migration equation.

(c) Annual rainfall is included as a proxy for better and worse growing years in traditional agriculture, lacking data on subsistence crop output.

(d) A typical tour of duty for a Mozambican miner is eighteen months. As in the case of Malawi, this means there is generally a lag in adjustment, and the number of mine workers lagged one year is therefore added to the migration equation.

The 1942 circular meant that "people were forced to work, but not forced to work for anyone in particular," and "[t]he effect of the abolition of the *prazos* [a system of indentured labor formally abolished by the 1930 Labour Code] was to make the search for labour competitive" (Vail and White, pp. 304 and 292). As an approximation, a competitive demand for labor may therefore be envisioned and is estimated in a simple Cobb-Douglas derived demand form with a flexible accelerator term. However, while forced labor remained common, the supply of labor to this market cannot be modeled by a standard supply of labor equation. Nonetheless, the plantations had to compete for their labor with the recruiters from abroad and with nonagricultural employers at home, both of whom offered legal alternative methods of voluntarily satisfying employment requirements, and this competition is likely to have been heightened after 1961 as forced labor diminished. Equations for labor demand and wage determination in the Mozambican plantation sector are therefore estimated as

$$(11)\quad lne_a = \varepsilon_0 + \varepsilon_1 lnw_a/p_a + \varepsilon_2 lna_g + \varepsilon_3 lne_{a(-1)}$$

$$lnw_a/p_a = \omega_0 + \omega_1 m + \omega_2 lnw_{na}/p_h + \omega_3 q_{cn} + \omega_4 fr$$

where p_a is price of plantation crops, a_g is the area of plantation actually used for growing, fr is a dummy set equal to one before 1962 for the period of forced registration, q_{cn} is construction output, and e_a, w_a, p_h, m, and w_{na} are as before. Plausible hypotheses with respect to the equations (11) are

$$(12)\quad \varepsilon_2, \varepsilon_3, \omega_1, \omega_2, \omega_3 > 0 \text{ and } \varepsilon_1,\ \omega_4 < 0.$$

C. *South Africa*

Since 1960, and predominantly since 1970, some 3.5 million black South Africans deemed not entitled to remain in White Areas have been forcibly relocated in an archipelago of exceedingly poor "homelands" or Black States. It is this group, combined with prior homeland dwellers, who are of concern here. Almost no black workers with Section 10 rights, who are legally entitled to remain in prescribed "White Areas" for more than 72 hours, work in the mines.

As with foreign labor, South African black miners are recruited on contract by TEBA and must return to their homelands between contracts. The main alternatives to mine labor are to help in own-account crop cultivation or to seek wage employment elsewhere in South Africa. With regard to the latter, however, two provisions of South African law warrant a modification in the formulation of (4). Jobs in industry pay considerably better than either mine or farm work, but a substantial fraction of these jobs are held by Section 10 workers to whom preference must be given in hiring. The second provision is that TEBA is banned from recruiting in prescribed "White Farm Areas." But this does not necessarily mean a worker from the homelands may not weigh the possibility of obtaining agricultural wage work rather than contracting as a miner, though preference in agricultural employment goes to those black laborers not yet forcibly removed from the White Farm Areas. The result of these provisions is that industrial and agricultural alternatives probably weigh quite differently in the considerations of potential miners, and in consequence these two expected wage terms and the mine wage are each inserted separately in the mine labor equation, the coefficients thus absorbing the unknown weights. However, since 1978, such deliberations have been largely irrelevant for South Africa's black population. The program of compulsory internment in the homelands has generated a massive pool of unemployment and TEBA has reported excess South African applicants after 1978. Estimation of the mine labor equation for South Africa is therefore confined to 1964 through 1978.

No attempt is made to model wage or employment determination in industry or agriculture in South Africa. Even unskilled blacks' wages in industry are far above those of miners. Real wages in agriculture have risen slightly in the 1970's, at a time when mine employment of South African blacks has risen dramatically, but the connection is probably spurious. The rise in farm wages has almost certainly been occasioned by the forcible removal of squatter families from White Farms on the grounds that they pose a security threat.

D. *Botswana and Lesotho*

Lastly, Botswana and Lesotho have sufficient in common with regard to specification to warrant joint discussion here. From both Botswana and Lesotho an excess supply of miners has existed since 1974. It is now difficult for a man without a Valid Reengagement Certificate, authenticating satisfactory prior performance, to obtain work and few novices are contracted (see my paper, 1985b). The mine labor equations for Botswana and Lesotho are therefore estimated from 1946 through 1973.

Two variants on the mine labor equations are estimated for both countries: one as in (4), the other replacing crop value by rainfall. In fact, the two countries are each subject to severe, periodic droughts, affecting both crops and livestock, and it seems reasonable to expect an increase in willingness to work in the mines at such times. The rainfall variant is particularly useful in the Lesotho context, however, for data on crop production are available only for occasional years during the colonial period, limiting the sample size and continuity for estimating (4).

In neither country is plantation farming significant, the Freehold Farms in Botswana being essentially cattle breeding and fattening stations employing comparatively little labor. Indeed, in the organized labor market of Botswana, the government plays a major role in setting minimum wages and hiring substantial amounts of labor, either directly or in parastatals paying well above any notion of opportunity cost (Michael Lipton, 1978). In Lesotho, wage employment re-

mains at a minuscule level.[16] No attempt is therefore made to model these labor markets and the domestic wage and employment are taken to be exogenous variables in both cases.[17]

III. Estimation and Results

The estimated equations for mine migration, crop production, cattle accumulation, and estate labor are presented in Tables 1–4, respectively. The t-statistics for a zero null hypothesis are given in parentheses beneath each coefficient estimate. The methods of estimation are indicated with each regression.[18] In addition to crop output and mine labor in each context, plus plantation wage and employment in Mozambique and Malawi, the mine wage is also treated as an endogenous variable for the three largest suppliers of mine workers: Malawi, Mozambique, and South Africa.[19] The Appendix Table gives units of measurement, means, and standard deviations of all variables in this paper.

[16]In earlier years, the South African government discouraged South African companies from opening plants in Lesotho, apparently partially to ensure a continued labor supply for the mines and White Farms.

[17]Annual data on wages and employment in Lesotho could not be located for the earlier years. The expected wage difference through the 1950's thus actually reflects variations in the mine wage around an intercensal trend.

[18]The following acronyms indicate the respective estimation techniques: OL = Ordinary least squares, IV = Instrumental variables, AR = Autoregressive correction by Beach-Mackinnon method; $ARF = AR$ method with Fair's extension of instrumental variables; and COF = Cochrane-Orcutt autoregressive correction for discontinuous samples, with Fair's extension of instrumental variables. For each equation in which a serial correlation correction is made, a common factor restriction test for the first-order autocorrelation specification is also undertaken. For purposes of this test nonlinear least squares is adopted with a likelihood ratio test if no simultaneity correction is involved, or the test suggested in Ronald Gallant and Dale Jorgenson (1979) for nonlinear instrumental variables. Unless otherwise mentioned, the first-order autocorrelation restriction could not be rejected on a 95 percent confidence test.

[19]Instrumental variables are provided by taking principal components of the exogenous variables for all of the labor-supplying countries plus wage rates of white miners in South Africa and prices of the various minerals obtained.

A. *Mine Labor Equations*

From each one of the major suppliers of labor, the number of miners has responded positively to the expected wage differential between mine and home when not constrained by immigration or emigration quotas.[20] Moreover, if the mine wage and expected home wage are included as separate arguments, then a hypothesis that the respective coefficients are equal and opposite in sign cannot be rejected on a 95 percent confidence test in any of the equations in Table 1, except in the case of Mozambique. Clearly, southern Africa has not been a labor reserve for the mining houses, in the sense of infinite elasticity of labor supply, and recognition of this has been the chief cause of centralized recruiting to exploit monopsony power as explored in my article (1985a).

In each country, more successful harvests or more ample rainfall have significantly reduced the incidence of mine work. Even from Botswana, Lesotho, and the South African homelands, in each of which crop-growing conditions are extremely difficult, better crops are seen within this simultaneous framework to have detracted from migration to the mines. And in Malawi, both enhanced subsistence crops and increases in the number of licensed tobacco growers have tended to diminish mine labor migration.

As population growth has swelled the various labor forces, mine employment has been promoted significantly from each country except Mozambique and South Africa, some of the employment needs of a growing labor force thus being met through emigration.[21] Indeed the reported coefficients on the logarithm of population exceed one, and with at least 95 percent confidence for Botswana and Lesotho, implying that a restricted specification with proportion of population as dependent variable would not have been appropriate.

[20]The only exception is for Lesotho in the semilogarithmic specification.

[21]In the equations for both Mozambique and South Africa, the coefficient on population proves insignificantly different from zero and this term is excluded from the results reported in Table 1.

TABLE 1—MINERS IN SOUTH AFRICA

	Botswana				Lesotho			
	Miners	Miners	*ln*(*m*)	*ln*(*m*)	Miners	Miners	*ln*(*m*)	*ln*(*m*)
Intercept	−10151	−19402	.897	−.588	−55927	−50462	−.581	−.699
	(1.14)	(2.68)	(0.61)	(0.53)	(6.26)	(4.42)	(1.18)	(0.67)
Expected	586	834	.026	.035	975	1010	.0015	.0028
Wage Diff.	(1.65)	(2.42)	(1.85)	(2.67)	(1.87)	(1.57)	(0.18)	(0.25)
Crop	−63.1		−.0022		−67.7		−.0014	
Value	(2.23)		(1.95)		(2.45)		(3.07)	
Rainfall		−5.64		−.247E-3		−16.9		−.309E-3
		(2.72)		(3.12)		(3.58)		(3.39)
Population	54.7	68.3			127	127		
	(5.18)	(8.27)			(23.0)	(11.6)		
ln(pop.)			1.42	1.65			1.73	1.76
			(6.61)	(10.15)			(22.60)	(10.84)
Est. method	*ARF*	*AR*	*ARF*	*AR*	*IV*	*AR*	*IV*	*AR*
No. obs.	26	27	26	27	16	28	16	28
Adj. R^2		.76		.99		.89		.99
SD dep. var.	3570		.514		23123		.350	
SE regr.	1834		.074		3138		.051	
D-W stat.					1.69		1.86	
Rho	.333	.378	.373	.408		.521		.459
T-stat *rho*	(1.69)	(2.01)	(1.90)	(2.21)		(3.11)		(2.60)

	Malawi		Mozambique		South Africa			
	Miners	*ln*(*m*)	Miners	*ln*(*m*)	Miners	Miners	*ln*(*m*)	*ln*(*m*)
Intercept	−61139	−16.8	−12357	1.96	44817	57840	1.56	1.11
	(1.75)	(1.84)	(0.44)	(0.80)	(1.01)	(1.64)	(1.31)	(0.93)
Expected	5350	.041	4480	.036				
Wage Diff.	(3.79)	(1.44)	(2.55)	(2.57)				
Crop	−5389	−.069			−2968	−2900	−.013	−.0096
Value	(3.10)	(1.67)			(2.78)	(2.79)	(2.76)	(1.82)
Rainfall			−24.1	−.206E-3				
			(1.53)	(1.63)				
Population	5.78							
	(0.73)							
ln(pop.)		2.56						
		(1.80)						
No. Tobac.	−.167	−.367E-5						
Growers	(1.25)	(1.16)						
Construct.			−15.1	−.129E-3				
Output			(1.34)	(1.44)				
Period			10766	.083				
Forced Reg.			(1.97)	(1.93)				
Mine Wage					4874	4944	.020	.019
					(8.62)	(9.11)	(7.57)	(7.23)
Expected					−3914	−4080	−.017	−.015
Wage Indus.					(4.75)	(5.48)	(4.36)	(4.05)
Expected					1842		.0082	
Wage Agric.					(0.47)		(0.45)	
Lagged	1.07	.579	.787	.808	.922	.912	.885	.920
Dep. Var.	(8.60)	(2.19)	(3.47)	(3.65)	(11.04)	(11.32)	(9.94)	(10.16)
Est. method	*COF*	*COF*	*ARF*	*ARF*	*ARF*	*ARF*	*ARF*	*ARF*
No. obs.	13	13	23	23	29	29	29	29
SD dep. var.	63581	1.20	25267	2.00	56176	55213	1.41	1.36
SE regr.	6423	.146	5685	.045	11622	11428	.053	.052
Rho	−.417	−.328	−.597	−.605	−.496	−.480	−.474	−.459
T-stat *rho*	(1.66)	(1.25)	(2.55)	(2.61)	(2.86)	(2.79)	(2.68)	(2.58)

From both Malawi and Mozambique, the longer tours of duty combined with the South African Master and Servants Law against desertion are seen to have significantly retarded responsiveness of labor to changing conditions, for the lagged endogenous variable terms are significantly greater than zero.[22] In conducting specification tests for inclusion of a lagged endogenous variable, a zero null hypothesis could not be rejected for Botswana or Lesotho, but is strongly rejected for South Africa and such a term is therefore included in Table 1 for the South African case.[23] Although South African black migrants are required by law to return to their homelands each year between contracts, as if they were foreign, it seems some continuity in contracts does occur.[24]

South African black labor has been attracted to the mines by rising mine wages and deterred from accepting such contracts when the likelihood of obtaining the far higher industrial wage has increased. However, a negative coefficient is not found on the expected wage in agriculture within South Africa: a hypothesis that changes in agricultural employment conditions have left mine recruiting in South Africa unaffected cannot be rejected, partly because of the ban on TEBA recruiting in White Farm Areas to protect the white farmers.

Even from Mozambique, miners have been able to respond to the differential in earnings. Moreover, as construction output has expanded, presumably employing additional labor, Mozambican workers have been drawn away from mine work, though not always voluntarily. Indeed, the ending of widespread forced labor in the Portuguese colony is seen to have significantly reduced the mine labor available to the Rand. Prior to that, even the harsh conditions in the mines could seem appealing relative to forced labor in Mozambique.

B. *Crop Production and Cattle Accumulation*

In looking at the impact of out-migration on rural areas, there has been much discussion of the consequences of labor withdrawal in reducing subsistence output. The estimated crop production responses in Table 2 reject a simple labor surplus theory; the negative coefficients estimated on the employment variables indicate that increased wage employment (at home or in the mines) from any given level of population has reduced traditional crop output, *ceteris paribus*.

But other things are not equal. The withdrawal of labor for wage employment is also seen in three out of four cases to have significantly raised productivity in crops through additions to the capitalized wage bill.[25] Moreover, in Table 3, the sizes of the cattle herds in Botswana and Malawi are also shown to have been significantly enhanced as wage earnings have accumulated, though the effect is statistically weaker for South Africa.[26] These findings corroborate the results of a growing number of case studies, which illustrate the importance of wage earners in the household to both peasant crop and animal husbandry (Carol

[22] In the case of Malawi, the estimated coefficient on the lagged endogenous variable slightly exceeds one in the linear specification, though not significantly so, and if the equation is reestimated with the first difference in mine labor as dependent variable, other coefficients remain unaffected.

[23] It should however be mentioned that for the second equation reported for Botswana, the common factor restriction test is rejected on a 95 percent confidence likelihood ratio test, though not at a 2.5 percent significance level.

[24] In diamond mining, South African black workers have gained access to more highly skilled jobs previously reserved for white miners and have tended to become more permanent workers with higher pay. In addition, so-called "colored" workers also tend to possess more permanent jobs.

[25] The capitalized wage and crop incomes are highly correlated for Malawi. Table 2 therefore reports variants in which each appears separately. Separate estimates are not reported for size of Malawi's cattle herd in Table 3, despite collinearity, for exclusion of the separate terms leaves the results unaffected. For all countries, the results for both crops and cattle prove quite insensitive to the choice of depreciation rate in computing the capital stocks. Estimates were obtained with stocks evaluated at 5, 10, and 20 percent, and the initial stocks accordingly adjusted.

[26] The marginal significance level on the common factor restriction test for the Botswana cattle equation is, however, particularly low at only 0.6 percent.

TABLE 2—CROP PRODUCTION
(Dependent variable: ln(output/pop.))

	Botswana		Lesotho		Malawi			South Africa
Intercept	−8.53	.487	34.6	5.07	−10.9	−8.89	−13.7	−12.9
	(1.16)	(0.22)	(1.20)	(0.70)	(2.89)	(4.42)	(4.75)	(2.13)
ln(accum.	1.79	1.80	−2.45		.474	.579		2.63
wages/pop.)	(2.18)	(2.15)	(2.90)		(1.18)	(3.20)		(2.12)
ln(accum.	.787		−5.31		.366		1.46	1.77
crops/pop.)	(1.28)		(1.50)		(0.35)		(3.18)	(2.66)
ln(employm.	−2.07	−3.43	−1.47	−3.50	−.546	−.691	−.769	−3.62
/pop.)	(1.30)	(2.80)	(0.61)	(1.92)	(1.27)	(1.55)	(1.90)	(2.44)
Rainfall	.0020	.0020	.0018	.0013				.343E-3
	(2.90)	(2.75)	(1.56)	(1.73)				(0.75)
Fertilizer					.975E-5	.105E-4	.165E-4	
					(1.27)	(1.60)	(3.07)	
No. Tobacco					.756E-5	.774E-5	.737E-5	
Growers			.(3.46)	(3.59	(3:33)			
ln(time)				.902				
				(2.53)				
Est. method	*IV*	*IV*	*IV*	*IV*	*ARF*	*ARF*	*ARF*	*IV*
No. obs.	31	31	19	19	31	31	31	31
SD dep. var.	.693	.693	.593	.593	.724	.735	.724	.385
SE regr.	.505	.516	.437	.442	.162	.162	.167	.233
D-W stat.	2.03	1.81	1.70	1.74				2.11
Rho					.544	.573	.544	
T-stat *rho*					(3.19)	(3.53)	(3.15)	

TABLE 3—SIZE OF CATTLE HERD
(Dependent variable: Number of cattle)

	Botswana	Malawi	South Africa
Intercept	−15198	4054	436
	(0.07)	(0.26)	(0.60)
Accumulated	8.37	.150	.140E-3
Wages	(2.73)	(1.65)	(1.26)
Accumulated	8.64	−.004	.005
Crops	(0.81)	(0.09)	(0.97)
Rainfall	167		.552
	(1.69)		(1.51)
Lagged	.644	.963	.613
Dep. Var.	(3.95)	(10.84)	(3.53)
Est. method	*AR*	*AR*	*OL*
No. obs.	30	27	28
Adj. R^2	.88	.99	.45
Durbin's *H*			−.98
Rho	.574	−.428	
T-stat *rho*	(3.69)	(2.23)	

Kerven, 1983). More generally, it appears that wage income of migrants can enhance the working or fixed capital available to the traditional, rural sector, raising productivity in the longer run while in the short-run reducing output.

To this, there is an exception in Table 2 for Lesotho.[27] Earlier in this century, Basutoland was a net exporter of grain. Today it is utterly dependent on emigrant labor; crop production has been neglected and massive soil erosion has ensued (James Cobbe, 1982). Whether the soil erosion would have occurred anyway, resulting from extensive deforestation as population has grown, migration notwithstanding, is unclear. Nonetheless, the second equation reported for Lesotho does indicate a rising trend in per capita crop production offset by the large negative effect of labor withdrawal for employment.[28]

[27]The results on accumulated wages and crop income for Lesotho are independent of whether these terms are included separately or even as a combined capital stock figure.

[28]An exponential time trend added to the crop equations for Botswana and South Africa proves insignificantly different from zero and does not change the estimated effects of other variables included in Table 2. In the case of Malawi, accumulated wages, accumulated crop income, and time are too highly correlated to disentangle successfully.

It may also be mentioned that evidence of accumulated crop income enhancing crop production is clearly significant only for South Africa and there is no evidence to suggest that investments in cattle occur out of crop income in any context.[29] It thus seems some degree of out-migration may even be a requirement of long-term development for traditional agriculture.[30]

C. *Estate Labor*

Almost all of the hypotheses listed with respect to estate labor in (10) and (12), for Malawi and Mozambique, respectively, are supported by the results in Table 4.[31] But the most important terms, from the present perspective, relate to mine labor.

In Mozambique, recruiting of labor for the Rand mines is shown to have made plantation labor significantly more expensive, and in turn higher wages resulted in significantly diminished employment. This raises the question of why the Portuguese colonial authorities would allow such recruiting, imposing increased costs on the European plantation owners, diminishing plantation employment and hence presumably reducing crop production and export. The answers are essentially twofold. (a) A gap existed between the authorities and plantation owners in regard to their concerns in running the colony. "[T]he specifically anti-capitalist and pro-nationalist ethic of the Estado Novo...predisposed the Administrators to attack the foreign-capitalized companies" (Vail and White, p. 299). (b) And in particular, Portugal managed to gain directly from South African mine recruitment through at least three provisions of various agreements negotiated with the South African state: (*i*) the South Africans agreed to ship specified minimum amounts through the port of Lourenco Marques in return for the right to recruit; (*ii*) a recruiting tax was collected by the Portuguese; (*iii*) but by far the most lucrative arrangement for the Portuguese was a secret clause whereby compulsory deferred pay, kept until the miners returned to Mozambique, was retained by Portugal in the form of gold at a premium price.

Table 4 also shows the major role played by forced labor in holding down plantation wages in Mozambique prior to 1962. The subsequent rise in wage was indeed associated with an absolute decline in plantation employment thereafter, with magnitudes such that the estimated coefficients in Table 4 indicate a rise in expected wage in agriculture, *ceteris paribus*. Such a rise would in itself serve to discourage mine migration, but, according to Table 1, the ending of forced labor reduced such migration beyond any effect on expected wage alone.

In Malawi, too, recruiting by TEBA is seen in Table 4 to have significantly diminished the supply of labor to the estates. Before independence, the colonial authorities did impose quotas on recruiting precisely to hold down such wages, though no doubt this was mitigated by the fact that Nyasaland and South Africa were both British, as protectorate and dominion territory, respectively. In the aftermath of independence, a Malawian Government was understandably reluctant to constrain its own

[29] Current crop income is an argument in computing capitalized crop income, and this might explain why a weak positive association is found between accumulated crop income and crop productivity but not herd size. However, if $ln(kq/N)$ is replaced by $ln(kq/N)_{-1}$ in the crop equations, the results are unaffected.

[30] In addition, fertilizer inputs and permits for tobacco growers have clearly raised smallholder crop production in Malawi and the importance of adequate rains to both crops and cattle in Botswana, to crops in Lesotho, and cattle in South Africa is apparent. It is not clear why the coefficient on rainfall tends to be negative if inserted into the crop or cattle equations for Malawi, but the inclusion of this term has no significant effect on other coefficients.

[31] The solitary exception is for the effect of the number of tobacco growers on estate labor supply in Malawi which is indistinguishable from zero. Two variants of the wage equation for Mozambique are displayed, with and without construction output included, since correlation between construction and the nonagricultural wage term affects the significance of the latter, though clearly the domestic employment alternative in construction or otherwise proves relevant either way. Adopting an *F*-specification test, the flexible accelerator terms appended to the demand equations for both Malawi and Mozambique prove significantly greater than zero in a 95 percent confidence test or better. However, no such term is included in the supply equation for Malawi since it proves indistinguishable from zero in a likelihood ratio test.

TABLE 4—ESTATE LABOR
(Dependent variable: ln(employment in agriculture))

Malawi			
Intercept	−.421	Intercept	−2.55
	(0.38)		(0.66)
ln(agric. wage /output price)	−.097	ln(agric. wage /cpi)	.660
	(1.41)		(1.28)
ln(area tea)	.279	ln(nonag. wage /cpi)	−.382
	(2.93)		(1.42)
ln(area tobacco)	.239	ln(mine labor)	−.114
	(3.09)		(2.08)
Lagged Dep. Var.	.577	ln(maize price /cpi)	−.105
	(4.21)		(1.13)
		ln(no. tobacco growers)	.048
			(0.42)
		ln(population)	1.75
			(3.73)
Est. method	*IV*		*ARF*
No. obs.	21		31
SD dep. var.	.225		.766
SE regr.	.052		.108
Durbin's *H*	.547		
Rho			.754
T-stat *rho*			(6.13)

Mozambique	ln(employment in agric.)			ln(real wage product)
Intercept	−1.32	Intercept	4.21	2.42
	(1.04)		(2.74)	(2.44)
ln(agric. wage /output price)	−.386	Mine labor	.611E-5	.105E-4
	(8.00)		(1.62)	(3.55)
ln(area plantations)	1.00	ln(nonag. wage /cpi)	.099	.314
	(7.41)		(0.48)	(2.16)
Lagged Dep. Var.	.218	Construction output	.404E-3	
	(2.21)		(1.47)	
		Period of forced registration	−.672	−.601
			(5.70)	(5.52)
Est. method	*IV*		*ARF*	*ARF*
No. obs.	21		21	21
SD dep. var.	.123		.864	.883
SE regr.	.046		.117	.121
Durbin's *H*	.632			
Rho			−.444	−.455
T-stat *rho*			(2.08)	(2.20)

people from electing to migrate to the mines. But as estates passed increasingly into Malawian hands, this strategy gathered opponents until recruiting was suspended entirely in 1974.[32] The result was almost an 80 percent rise in estate employment by 1978, and a 15 percent drop in the real wage of estate workers in the first two years after the cutoff. But the additional employment on estates amounted to exactly half the decline in numbers previously at the Rand mines and the slight increase in nonagricultural employment accounted for less than 9 percent extra. The Customary Lands had to

[32] Evidence exists indicating that suspension was planned even prior to the plane crash in April 1974. See Robert Christiansen and Jonathan Kydd (1983).

absorb many of the displaced workers. In general, Table 4 shows the estates have had to compete for their labor against prices of crops grown on these Customary Lands as well as with nonagricultural employers. Yet with this massive infusion of labor, ADMARC (the state marketing board) could increase its intake while employment rose and wages fell on the estates. From the Customary Lands, ADMARC purchases of tobacco, cotton, maize, groundnuts, and coffee were nearly 60 percent higher in 1978 than in 1974 (measured in constant 1974 prices). Tobacco purchases alone more than doubled, in part because the number of licensed growers on Customary Land was permitted to rise sharply. Nonetheless, discontent resulted and the ban on TEBA recruiting was lifted in June 1977, but by then South Africa no longer trusted Malawi as a source and the number of mine workers from Malawi has never recovered (Robert Christiansen and Jonathan Kydd, 1983).

IV. Closing Remarks

Two major themes pervade this paper. The first extends the analysis of labor withdrawal from agriculture to embrace long-run effects. The development literature on surplus labor has focused on the marginal product of labor either *ceteris paribus* or *mutatis mutandis*, after labor input adjustments by residual rural dwellers (Amartya Sen, 1966). What has been neglected is the possibility that earnings of migrants may serve as a source of capital accumulation in the rural areas.

By specifying a simultaneous framework in which declining agricultural output may increase the propensity to migrate, but also labor departure may diminish agricultural output in the short run while enhancing productivity in the long run, it has been possible to examine these effects separately. Thus, emigration to the South African mines has been shown to have reduced crop production in the subsistence sectors of Botswana, Lesotho, Malawi, and the South African homelands in the short run. But the results also suggest that earnings of migrants have enhanced both crop productivity and cattle accumulation in the longer run, except in Lesotho. Whether the mechanism of productivity enhancement is one of physical investment, financing of new techniques, or insurance permitting experimentation with riskier methods cannot be discerned. Given incomplete insurance markets and segmented capital markets, each of these mechanisms may be important, and may effectively serve to lower the shadow cost of labor withdrawn from agriculture.

The other major theme addresses domestic wage formation, desire to emigrate and limits imposed on the flow of migrants. In the context of each of the five regions examined, the gap between the wage available in the South African mines and the domestic wage weighted by probability of employment, is estimated positively to affect the desire to be a miner. In Botswana and Lesotho, this is assumed not to have had any feedback effect on domestic wage and employment, because such wage jobs as exist are largely in the public sector. In Malawi, Mozambique, and South Africa, the story is different.

In both Malawi and Mozambique, emigration to South Africa's mines has significantly inflated labor costs to the local estate and plantation operators. Since, unlike most internal migration, international migration may be fairly readily subjected to control, this raises a fascinating conflict in terms of political reaction. In colonial Nyasaland, the British imposed quotas on the numbers permitted to be recruited for the mines, precisely to protect the white estate owners. After independence, the Malawian government initially took a broader view. However, the resulting surge in recruiting soon placed pressure on the now Malawian estate owners and ultimately recruiting was suspended entirely, with some short-term improvement in smallholder production and drop in estate wages. Yet domestic employment remained inadequate to absorb the restricted miners and Malawi tried subsequently, though with limited success, to increase again the recruiting activity. In Mozambique, the Portuguese authorities struck an extremely lucrative deal with South Africa as a price for recruiting privileges. The European plantation owners in Mozambique were to some extent compensated for this by the fact that the authori-

ties implicitly permitted continued use of forced labor. Within South Africa itself, the government faced both the interests of the white farmers and of the mining houses. Recruiting for the mines is banned in the White Farm Areas, so no direct pressure on agricultural wages has been permitted. But the political dominance of the Afrikaans farmers is on the wane. Some 3.5 million black South Africans have been forcibly relocated in the Black States, many being removed from White Farm Areas. The pool of South African labor potentially available to the mines has thus dramatically swelled. And the mining houses have swung their recruiting toward South African workers rather than rely on foreign labor supply, which has become inherently unstable with the Malawian suspension and independence in Mozambique.

APPENDIX TABLE—SAMPLE MEANS[a]

		Botswana	Lesotho	Malawi	Mozambique	South Africa
Local Currency Units (LCU)		Rands[b]	Rands[b]	Kwacha	Escudos	Rands
m	No. miners	22104 (5506)	65466 (19589)	38237 (43315)	125676 (7667)	239943 (33112)
$v_m - \frac{w_h}{p_h}\frac{e}{n-m}$	'70 Rnds/mth.	10.26 (1.95)	16.41 (1.90)	12.08 (2.17)	14.27 (1.59)	–
$\frac{p}{p_h}\frac{q}{n-m}$	'70 LCU/year	36.88 (21.84)	83.37 (40.50)	1.69 (1.49)	–	9.71 (4.75)
q	Index '70 = 1.0	4.769 (2.397)	1.035 (.378)	.873 (.424)	–	.936 (.254)
c	No. cattle	1431700 (514787)	–	399407 (120343)	–	3642 (239)
r	Mm./year	508 (156)	708 (140)	–	1026 (145)	823 (116)
kw	1000 '70 LCU	75239 (44779)	143138 (54254)	311540 (186508)	–	2082590 (1026307)
kq	1000 '70 LCU	28675 (5927)	91055 (7608)	686976 (207760)	–	158454 (22848)
cw	1000 '70 LCU	43370 (26617)	–	179346 (99546)	–	1253217 (662018)
cq	1000 '70 LCU	13446 (4010)	–	383803 (109241)	–	75786 (13227)
N	1000 people	522 (89)	875 (138)	3823 (744)	–	18272 (4198)
e	No. employed	23072 (13659)	2653 (2044)	171203 (55055)	–	–
e_a	No. employed	–	–	63444 (25318)	110656 (15136)	780062 (49459)
w_a/p_a	'70 LCU/mth.	–	–	9.53 (2.28)	203 (84)	–
w_a/p_h	'70 LUC/mth.	–	–	7.42 (1.44)	–	9.07 (1.50)
w_{na}/p_h	'70 LCU/mth.	–	–	25.40 (8.67)	514 (180)	43.67 (8.82)
Fertilizer	Short tons	–	–	9484 (13501)	–	–
p_c/p_h	'70 Kw/sh. ton	–	–	30.20 (8.11)	–	–
n_{to}	No. growers	–	–	71541 (15661)	–	–
a_{to}	Acres	–	–	25221 (8713)	–	–
a_{te}	Acres	–	–	33401 (5147)	–	–

APPENDIX TABLE—(CONTINUED)

		Botswana	Lesotho	Malawi	Mozambique	South Africa
Local Currency Units (LCU)		Rands[b]	Rands[b]	Kwacha	Escudos	Rands
a_g	Hectares	–	–	–	239646 (32767)	–
q_{cn}	Mill. '70 Esc	–	–	–	381 (168)	–
e_{na}	No. employed	–	–	–	–	845621 (350422)
v_m	'70 Rnds/mth.	–	–	–	–	19.33 (8.01)

[a]Sample sizes are consistent with the equation in which each variable appears. Standard deviations are shown in parentheses.
[b]The Pula and Maloti were introduced into Botswana and Lesotho, respectively, after the termination of the sample period.

REFERENCES

Boeder, Robert B., "Malawians Abroad," unpublished doctoral dissertation, Michigan State University-East Lansing, 1974.

Breytenbach, W. J., *Migratory Labour Arrangements in Southern Africa*, Pretoria: Africa Institute of South Africa, 1979.

Byerlee, Derek, "Research on Migration in Africa," Africa Rural Employment Paper No. 2, Michigan State University-East Lansing, September 1972.

Christiansen, Robert E. and Kydd, Jonathan G., "The Return of Malawian Labour from South Africa and Zimbabwe," *Journal of Modern African Studies*, October 1983, *21*, 311–26.

Cobbe, James, "Labour-related Aspects of Rural Development in Lesotho," Institute of Labour Studies, Discussion Paper No. 6, National University of Lesotho, Maseru, 1982.

Collier, Paul and Lal, Deepak, "Poverty and Growth in Kenya," *World Bank Staff Paper*, No. 389, May 1980.

Dean, Edwin, *The Supply Responses of African Farmers*, Amsterdam: North-Holland, 1966.

de Vletter, Fion, "Conditions Affecting Black Migrant Workers in South Africa," in W. Roger Bohning, ed., *Black Migration to South Africa*, Geneva: ILO, 1981.

Duffy, James, *A Question of Slavery*, Cambridge: Harvard University Press, 1967.

Elkan, Walter, "Labour Migration from Botswana, Lesotho and Swaziland," *Economic Development and Cultural Change*, April 1980, *28*, 583–96.

First, Ruth, *Black Gold: The Mozambican Miner, Proletarian and Peasant*, New York: St. Martin's Press, 1983.

Gallant, A. Ronald and Jorgenson, Dale W., "Statistical Inference for a System of Simultaneous, Non-linear, Implicit Equations in the Context of Instrumental Variable Estimation," *Journal of Econometrics*, October/December 1979, *11*, 275–302.

Gordon, Elizabeth, "Easing the Plight of Migrant Workers' Families in Lesotho," in W. R. Bohning, ed., *Black Migration to South Africa*, Geneva: ILO, 1981.

Greenwood, Michael J., "Research on Internal Migration in the United States," *Journal of Economic Literature*, June 1975, *13*, 397–433.

Halpern, J., *South Africa's Hostages*, Baltimore: Penguin, 1965.

Harris, John R. and Todaro, Michael P., "Migration, Unemployment and Development," *American Economic Review*, March 1970, *60*, 126–42.

Kerven, Carol, "The Impact of Wage Labor and Migration on Livestock and Crop Production in African Farming Systems," paper presented at the Third Annual Farming Systems Symposium, Kansas State University, October-November 1983.

Kowett, Donald K., *Land, Labour Migration*

and Politics in Southern Africa, Uppsala: Scandinavian Institute of African Studies, 1978.

Lipton, Merle, "Men of Two Worlds," *Optima*, November 1980, *29*, 72–201.

Lipton, Michael, *Employment and Labour Use in Botswana*, Gaborene: Ministry of Finance and Development Planning, 1978.

Lucas, Robert E. B., "The Supply-of-Immigrants Function and Taxation of Immigrants' Incomes," in Jagdish N. Bhagwati, ed., *The Brain Drain and Taxation*, Amsterdam: North-Holland, 1976.

______, (1985a) "Mines and Migrants in South Africa," *American Economic Review*, December 1985, *75*, 1094–108.

______, (1985b) "Migration Amongst the Batswana," *Economic Journal*, June 1985, *95*, 358–82.

______, (1985c) "The Distribution and Efficiency of Crop Production in Tribal Areas of Botswana," *World Bank Staff Working Papers*, No. 715, 1985.

______ **and Stark, Oded,** "Motivations to Remit: Evidence from Botswana," *Journal of Political Economy*, October 1985, *93*, 901–18.

Murray, Colin, *Families Divided: The Impact of Migrant Labour in Lesotho*, London: Cambridge University Press, 1981.

Penvenne, Jeanne, "Forced Labor and the Origin of an African Working Class: Lourenco Marques, 1870–1962," Working Papers in African Studies No. 13, African Studies Center, Boston University, 1979.

Psacharopoulos, George, "Estimating Some Key Parameters in the Brain Drain Taxation Model," in J. N. Bhagwati, ed., *The Brain Drain and Taxation*, Amsterdam: North-Holland, 1976.

Rempel, Henry and Lobdell, Richard A., "The Role of Rural-to-Urban Remittances in Rural Development," *Journal of Development Studies*, April 1978, *14*, 324–41.

Schapera, Isaac, *Migrant Labour and Tribal Life*, London: Oxford University Press, 1947.

Schultz, T. Paul, "Notes on the Estimation of Migration Decision Functions," in Richard H. Sabot, ed., *Migration and the Labor Market in Developing Countries*, Boulder: Westview Press, 1982.

Sen, Amartya K., "Peasants and Dualism with or without Surplus Labor," *Journal of Political Economy*, October 1966, *74*, 425–50.

Stahl, Charles W. and Bohning, W. Roger, "Reducing Dependence on Migration in Southern Africa," in W. R. Bohning, ed., *Black Migration to South Africa*, Geneva: ILO, 1981.

Stark, Oded, *Economic-Demographic Interaction in Agricultural Development*, Rome: F. A. O., 1978.

______, "On the Role of Urban-to-Rural Remittances in Rural Development," *Journal of Development Studies*, April 1980, *16*, 369–74.

Todaro, Michael P., *Internal Migration in Developing Countries*, Geneva: ILO, 1976.

Vail, Leroy and White, Landeg, *Capitalism and Colonialism in Mozambique*, Minneapolis: University of Minnesota Press, 1980.

Wilson, Francis, "International Migration in Southern Africa," *International Migration Review*, Winter 1976, *10*, 451–88.

International Labour Organization, *Report Concerning the Observance by the Government of Portugal of the Abolition of Forced Labour Convention*, Geneva, 1962.

A General Equilibrium Analysis of Partial-Equilibrium Welfare Measures: The Case of Climate Change

By MARY F. KOKOSKI AND V. KERRY SMITH*

This paper uses computable general equilibrium models to demonstrate that partial-equilibrium welfare measures can offer reasonable approximations of the true welfare changes for large exogenous changes. With consistency in the size and direction of the indirect price effects associated with large shocks, single-sector partial-equilibrium measures will exhibit small errors. Otherwise the errors can be substantial and difficult to sign.

The purpose of this paper is to evaluate the errors in partial-equilibrium measures of the welfare changes resulting from large multisector, exogenous shocks to an economic system. It would not be surprising to any economist to find that partial-equilibrium welfare measures would perform poorly in such cases. However, general answers to the questions of what is a large change, or how many sectors need to be involved before partial-equilibrium methods break down, are not available. We begin the process of developing answers by illustrating a new use for computable general equilibrium (CGE) models—one that has been largely overlooked in the extensive literature that uses these models to evaluate the implications of domestic tax or international trade (for example, tariff) changes.[1] More specifically, we impose exogenous changes on the unit costs of producing commodities in several sectors with a CGE model of a developed economy, and compare partial-equilibrium welfare measures of their impact with the "true" general equilibrium measures. Our application also has independent interest. It is the first general equilibrium evaluation of the economic effects of a carbon dioxide (CO_2) induced climate change.

Based on estimates of the atmospheric concentration of CO_2 around the start of the industrial revolution in comparison with the 1983 measurements, CO_2 levels have increased approximately 22 percent. The CO_2 absorbs long-wave terrestrial radiation. This leads to increased resistance to the upward radiative transfer of heat and increases in the mean global temperature—the so-called "greenhouse effect." With a doubling in atmospheric CO_2, mean temperatures are expected to rise 1°C to 3°C, with greater increases at the poles. Melting of the polar ice caps and uncovering of land or ocean are expected to change wind and precipitation patterns. Our analysis focuses on scenarios designed to represent a 50 percent increase in CO_2 (over the industrial revolution levels). This corresponds to an increase likely to arise in under fifty years.[2]

*Economist, Division of Price and Index Number Research, U.S. Bureau of Labor Statistics, Washington, D.C. 20212, and Centennial Professor of Economics, Vanderbilt University, Nashville, TN 37235, respectively. Support for Kokoski's research was provided by the Lurcy Fellowship at the University of North Carolina-Chapel Hill, and we both acknowledge the Climate Dynamic Program of the National Science Foundation, grant nos. ATM 8217303 and ATM 8317619. Thanks are due John Shoven and John Whalley for assistance in locating related literature, Knox Lovell and three anonymous referees for constructive comments on earlier drafts of this paper, and to participants at the Microeconomics Workshop at Vanderbilt University, the NBER Applied General Equilibrium Workshop at Stanford University, and the Economics Workshop at Oak Ridge National Laboratory for comments.

[1]For a detailed overview of the results of these studies, see John Shoven and John Whalley (1984).

[2]For summary of the evidence on the accumulation of CO_2 and the climatic effects of CO_2 and other trace gases, see F. Kenneth Hare (1985, pp. 52–59) and John Firor and Paul Portney (1982, pp. 182–99).

Our results can be used to address two issues—partial-equilibrium welfare measurement and the potential economic impacts of climate change. On the first of these, the findings suggest that fairly large single-sector impacts (with as large as a 42 percent unit cost increase in one sector) can be adequately measured using a single-market partial-equilibrium measure of compensating variation. However, smaller multisector changes (in terms of the unit cost increases implied for each sector) exhibit large errors in single market, partial-equilibrium (*PE*) welfare measures. Moreover, none of our attempts to extend the *PE* measures to include several markets either "vertically" or "horizontally" was consistently superior to the other alternatives.

The results for the second issue—the economic impacts of CO_2 increase—are intended to be illustrative only. Our model is a small approximation of a developed economy, parameterized with U.S. data. The scenarios are also approximate descriptions of the types of impacts thought to arise from CO_2 induced climatic change. The specific findings depend on the effects specified to arise in sectoral production patterns as part of the scenario design. Two interesting features were consistently observed for all the scenarios we considered. The scenarios used to represent a CO_2 induced climate change can produce large and mixed price (and output) effects with both increases and decreases. Thus, partial appraisals of selected sectors can give a misleading view of the full economic impacts of such changes. Equally important, the policies did not uniformly affect low- and high-income households.

I. Background and Outline of the Analysis

A. *Background*

Two sets of research have considered problems related to general equilibrium welfare measurement. The first involves descriptions of the rules a government should use when evaluating the efficiency gains from public projects in a tax distorted economy. Much of this work is summarized in W. Erwin Diewert (1983). All of the analyses have tended to accept a Harberger-type measure of efficiency gains or losses,[3] and consider only differential changes resulting from a proposed policy action (usually defined as a tax change). They ignore indirect price and income changes from other markets on the market affected by the policies under evaluation. Robin Boadway's 1975 synthesis of Arnold Harberger's 1971 general recommendations for applied welfare analysis together with his specific suggestions for the factor prices used in public projects, given distorted markets (Harberger, 1969), and the Diamond-Mirrlees (see Peter Diamond and James Mirrlees, 1971a,b) arguments for the use of producer prices, all highlight this assumption.

The second set of research is more closely related to our objectives. It consists primarily of two papers. The first (John Whalley, 1975) evaluated simple, Harberger-type, partial-equilibrium (*PE*), extended partial-equilibrium (*EPE*), and general equilibrium (*GE*) measures of the efficiency implications of the removal of the distortionary U.K. capital income taxation system (with total tax receipts held constant). Whalley's analy-

[3]Diewert (1985) has recently compared two alternative measures of the efficiency losses from tax distortions in a general equilibrium context. They include a quantity index and price index. The quantity measure holds each individual at his initial, tax distorted utility level, and considers the number of multiples of the reference (i.e., associated with the tax distorted equilibrium) vector the economy can produce without these distortions. He attributes this measure to Maurice Allais (1943) and Gerard Debreu (1951). The price measure of efficiency loss is (minus) the sum of the Hicksian equivalent variations over all consumers with the Pareto optimal prices as the reference point. He also proposed a differential approach to sensitivity analysis in general equilibrium models. This approach is implemented by differentiating the loss measure with respect to second-order parameters characterizing either demand or supply features of the economy. His analysis suggests that the Allais-Debreu measure of inefficiency increases with increased substitution on either the demand or supply side of the economy. While this is an interesting alternative to the use of numerical methods in gauging the sensitivity of general equilibrium results to alternative parameterizations, it is not relevant to our objectives because we seek to evaluate the performance of approximations when aspects of the general equilibrium solution have been ignored.

sis used a CGE model of the U.K. economy and three estimates of the change in the value of the total product to gauge these efficiency gains. His *PE* scheme holds all prices but capital constant and used linearized marginal product schedules for capital in each sector together with an iterative price adjustment scheme to estimate the common, post-tax-removal price for capital. The sum of the changes in the areas under marginal product schedules (for capital in each sector) was used to measure the change in value of aggregate output. The *EPE* collapses the economy to a two-commodity framework and used a simple capital allocation and price adjustment procedure to measure the changes in the value of output. The last measure calculates it using his nine-sector, seven-household class CGE model for the U.K. economy. All three approaches were used in several scenarios varying the assumptions on the production elasticities of substitution. The results uniformly suggest that the *PE* and *EPE* measures were unreliable approximations of the true changes in the value of output.

The second study of this type by Lee Edlefsen (1983) confirmed Whalley's findings. He found that the Harberger deadweight loss triangle can overstate welfare losses above a general equilibrium equivalent variation measure by a factor of 2 to 4. Moreover, his results suggested that the majority of this error can be due to the indirect effects. They were based on a four-equation demand and supply model with functions specified to be linear in price ratios (i.e., three commodities were assumed to capture general equilibrium effects). From his description it appears that the model does not reflect the influence of price changes on income by revaluing endowments. To capture all these effects requires a CGE model that takes account of commodity and factor markets, including the roles of their respective prices for the value of consumers' endowments.

B. *Specification of the CGE Model*

To evaluate the performance of partial-equilibrium welfare measures, a fairly simple nine-commodity CGE model was designed to represent a developed economy. Parameter values for the model were derived from U.S. data sources. Because it is a small-scale model, it was not intended to be capable of offering policy insights for the U.S. economy. Rather, the use of actual data for the prototype economic structure provides one strategy for dealing with the parameterization issues raised in judging results obtained from CGE models. Shoven and Whalley (1984) acknowledge this issue, noting that the results of CGE models are quite sensitive to model parameterization, especially the selection of elasticity values. Unfortunately, systematic sensitivity analyses with simple models (see Glenn Harrison and Lawrence Kimbell, 1983) have not uncovered general conclusions. Our approach limits the set of feasible parameter values to those estimated for an existing economic structure.

The nine-commodity economy includes: labor, land, capital, energy, chemicals, consumer durables, construction, services, and agriculture. These commodity definitions were the result of a compromise strategy. The model sought to identify important sectors in the economy and identify sectors likely to be affected by a CO_2 induced climate change. The model includes three consumers (with each intended to reflect specific groups of households). Two are domestic and one specified to represent the foreign sector. The domestic households were differentiated by income. Consumers were endowed with land, labor, and capital. Labor was specified so that a work-leisure choice is a part of household decisions.

Each of the produced commodities in each model was assumed to be derived from a nested Cobb-Douglas CES cost function with constant returns to scale as in

$$(1)\qquad C_j = \prod_{k=1}^{N_j}\left[\sum_{i=1}^{m}\left(a_{jik}P_i\right)^{r_{kj}}\right]^{\alpha_{kj}/r_{kj}}$$

where C_j = average cost for the jth produced commodity;

P_i = price of the ith factor input;

N_j = number of CES subfunctions (i.e., aggregate factors) in the production process for the jth commodity;

m = potential number of factor inputs (8 in principle, if all commodities served as inputs and final goods);

a_{jik}, α_{kj}, r_{kj} = production parameters (with $\Sigma_k \alpha_{kj} = 1$ and $\Sigma_i a_{jik} = 1$).

The domestic household's utility functions were specified to follow a Stone-Geary form as in (2). The foreign sector had a Cobb-Douglas form:

$$U_t = \prod_{i=1}^{m+1} (Q_{it} - \gamma_{it})^{\beta_{it}} \tag{2}$$

with $\beta_{it} > 0$, $\sum_i \beta_{it} = 1$, $Q_{it} > \gamma_{it}$,

where U_t = utility level for household t;

Q_{it} = consumption of commodity i by household t;

β_{it} = marginal budget share for commodity i by household t;

γ_{it} = threshold consumption level of commodity i by household t.

Two aspects of our specification of the domestic household utility functions are potentially important to the welfare analysis. In contrast to the frequently used CES formulations, income elasticities will not be equal (and unity) across commodities. Equally important, the definition of the compensating variation (CV) must be adjusted from conventional expressions with this utility function (see G. W. McKenzie, 1983, p. 40) to reflect the effects of price changes on income changes resulting from changes in the value of household endowments as well as on earnings through revisions in the labor-leisure choice (see our 1985b paper for discussion of the implications of these adjustments), as in

$$CV = (R-1)\overline{M}^0 + \sum_i \left(P_i^1 - RP_i^0\right)\gamma_i + \overline{Q}_k\left(RP_k^0 - P_k^1\right) - d\overline{M}, \tag{3}$$

where $\overline{M}$ = income associated with an individual's endowments, where there is no "own consumption demand"; the superscript 0 designates the initial value;

$d\overline{M}$ = the change in this exogenous income as a result of the change in the prices of these endowments;

$\overline{Q}_k$ = endowment of labor time (Q_k);

$R = \prod_i (P_i^1/P_i^0)^{\beta_i}$.

C. *Parameterization of the Model*

The year 1972 was the base year for our characterization of consumer demand and the distribution of income between the two domestic households. The parameters for the utility functions were taken from estimates reported by D. Eastwood and J. Craven (1981) for the extended linear expenditure system, with two adjustments. Their model did not include a labor/leisure choice so we assumed (following Shoven and Whalley, 1972) that leisure comprised 3/7 of total labor time and calculated the budget share for leisure as 3/7 of the share of labor income. The *LES* subsistence parameters for all commodities but leisure were adjusted for each of the two domestic households using estimates in C. Lluch et al. (1977) to reflect the effect of income level. The low-income consumers' source of income and share of national income were specified based on the lower three quintiles and the higher that of the top two quintiles. The third consumer group represents the foreign sector with foreign sector income defined as the sum of exports to the United States in 1972. The shares of imports of total expenditures on these commodities determined the parameters of the function.

The parameters for the sectoral cost functions were obtained from a variety of sources —cost shares primarily from Census of Manufacturers, Mining or Construction in 1972. Substitution elasticities were selected based on examination of a range of studies, including the detailed results reported in Michael Hazilla and Raymond Kopp (1982). The specification of the cost function allowed some flexibility in assignment of substitution elasticities by grouping factors into separate subfunctions. Panel A of Table 1 reports the composition of these subfunctions for each of the six production sectors in the model by indicating the Allen gross elasticities of substitution for the pairs of inputs in each, and by specifying the nonzero input cost shares in the five rows below this.

TABLE 1—PRODUCTION AND DEMAND PARAMETERIZATION FOR THE CGE MODEL

	Sector/Commodity					
	Energy	Durables	Agriculture	Chemicals	Construction	Services
A. **Production**						
1) Gross Substitution Elasticity[a]						
I	–	–	–	–	$\sigma_{12}=.50$	$\sigma_{13}=.25$
II	$\sigma_{23}=2.0$	$\sigma_{13}=.10$	$\sigma_{14}=\sigma_{15}=$ $\sigma_{15}=-2.0$	$\sigma_{34}=2.0$	$\sigma_{34}=-1.0$	–
2) Cost Shares						
Labor	.15	.33	.15	.25	.42	.69
	(.16)	(.42)	(.14)	(.27)	(.36)	(.60)
Land	.20	–	.80	–	.38	–
	(.10)		(.80)		(.43)	
Capital	.65	.57	.05	.38	.20	.11
	(.74)	(.50)	(.03)	(.67)	(.14)	(.20)
Energy	–	.10	.00	.38	.00	.20
		(.08)	(.01)	(.06)	(.07)	(.20)
Chemicals	–	–	.00	–	–	–
			(.02)			
3) Capital-Labor Ratio						
Benchmark	4.63	1.19	.21	2.48	.45	.33
Base Case Solution	4.36	1.70	.33	1.47	.47	.16
B. **Demand Elasticities**						
1) Eastwood-Craven Own-Price[b]	−.49	−.82	−.23	−.73	−.65	−.54
2) Base Case Solution Own-Price						
Low Income	−.951	−.957	−.641	−.960	−.705	−.865
High Income	−.923	−.933	−.539	−.972	−.763	−.972
Income						
Low Income	1.157	1.159	.734	1.183	.787	1.034
High Income	1.060,	.992	.594	1.135	.832	1.126

[a] s_i = Share of total costs associated with ith factor. Subscript i corresponds to the row labels for commodities in each model (i.e., developed and less developed).

σ = Gross (holding output constant) elasticity of substitution. In terms of parameters of equation (1) $\sigma=[r_k-(1-\alpha_k)]/\alpha_k$, where k corresponds to the kth CES nest. Both factors must be members of the same CES nest for this equation to be valid. $(1-r_k)$ designates the net elasticity of substitution for any pair of factors in the k nest.

[b] These elasticities correspond to estimates reported for closely related commodity categories. For a discussion of our adjustments to derive the values for our definitions, see Kokoski (1984).

Following the substitution elasticities we report some information on how well the base case solution reproduced the initial features of the data used to parameterize the production side of the model. These results will be affected by the interaction of the production and the demand specifications. Consequently, they need not reproduce the values of the data used to parameterize each component of the model.[4] Our summary includes both the cost shares for factor inputs and the capital-labor ratios from the benchmark data versus those implied by the model's base case. For example, the energy sector was specified to have three inputs—labor, land (to designate primary materials inputs), and capital. Capital and land were

[4] Most CGE models have been parameterized by calibrating the models (see A. Mansur and Whalley, 1984). That is, a subset of the parameters are set so the model reproduces a base case expenditure pattern. As Diewert (1985) observed, this procedure usually implies that preferences and technology sets are reasonably well approximated to the first-order but not to the second-order. In the past, use of restrictive functional forms, such as the CES, to characterize both preferences and technology have limited the distortions to second-order properties induced by these calibration techniques.

included in one subfunction and labor in another. The cost shares implied by the base solution are given first and then those used in the parameterization are shown below in parentheses. Thus, labor was specified to account for 16 percent of total costs of energy production while the base case solution implied 15 percent.

Panel B of Table 1 reports some comparable information for the demand side of the model. The approximate own-price elasticities from Eastwood and Craven are compared with the calculated values at the base case for the low- and high-income households and the base case calculated values of the income elasticities. In this case we would not expect close comparability because the Eastwood-Craven study did not provide estimates for households at different income levels and it did not incorporate a labor/leisure choice. Our estimates for the base solution reflect the adjustments made. Nonetheless, the overall results indicate a reasonably good correspondence for both sides of the economy. Of course, the match need not be exact since our objectives require only an economically plausible description of a developed economy.

D. *Scenario Design*

A CO_2 induced climate change will have at least two important features—a temperature increase and a precipitation change. The direction of the change in precipitation has not been as clear as the effect on temperature. Our scenarios have been designed recognizing the range of possibilities for these physical changes and using the available information on the effects of climate by sector to postulate the percentage changes in unit costs of production for each of a set of sectors.

It is reasonable to expect that the climate changes associated with a 50 percent increase in atmospheric CO_2 will impact several sectors simultaneously. Nonetheless, to provide information to judge the quality of partial-equilibrium welfare measures, we have specified single- and multiple-sector impacts of varying sizes. The single-sector scenarios involve effects only on agriculture while the multiple-sector effects involve from two to four sectors simultaneously. Our overall analysis has considered a large number of scenarios (and several specifications for the CGE economies).[5] However, we have limited the results presented here to seven cases.

The connection between a postulated change in temperature and precipitation and sectoral costs is probably best for agriculture. Our scenarios should be treated as judgmental summaries of the estimated effects from past studies of similar hypothesized climate changes. For example, using the case of agriculture, several studies have considered crop and area-specific analyses of changes comparable to those implied by our smallest impact cases. Early analyses of a 1°C temperature increase and 10 percent decline in precipitation implied a 26 percent reduction in the yields of corn (Wilfred Bach et al., 1981), 10 percent decline in wheat yields, and 26 percent decline in soybeans (Louis Thompson, 1975) for a weighted (by share of U.S. crop production) decline of about 22 percent. More recent evidence discussed by Paul Waggoner in the National Academy of Sciences report *Changing Climate* (1983) indicates two sets of estimates —one based on regression yield models and a second on a simulation of yield changes due to climate. The first implies the same climate change from about 2 to 13 percent decline for these same crops (with a weighted average close to 10 percent). By contrast, the second has a larger effect—about a 24.5 percent decline. We selected estimates at the higher end of this range (22 percent) to indicate the maximum impact and ignored adjustments in input mix or technology that might reduce the effect.

Table 2 defines the seven scenarios we report. In each case we report the features of the postulated climate change that provides the basis for selecting the unit cost impacts. The results in this table and in Table 3 provide a basis for gauging the implications of these changes for economic activities com-

[5]See Kokoski (1984) and our paper (1985a) for a more detailed discussion of some of this work.

TABLE 2—CLIMATE IMPACT SCENARIOS AND THEIR EFFECTS ON PRICES

Scenario Design	Climate Parameters	Increase in Unit Costs (percent)	Percent Change in Commodity Prices: Labor	Land	Capital	Energy	Chemicals	Durables	Const.	Services	Agric.
A. Single Sector											
1) Agriculture	1^0, +10%p	+4	−.5	+.2	−.5	−.5	−.5	−.5	−.3	−.3	+4.0
2) Agriculture	1^0, −10%p	+22	−3.0	+1.1	−2.8	−2.1	−2.4	−2.8	−1.4	−2.7	+22.0
3) Agriculture	2^0, −20%p	+42	−5.7	+2.2	−5.3	−4.1	−4.8	−5.3	−2.7	−5.1	+42.0
B. Multiple Sector											
4) Agriculture	1^0, −10%p	+22	−5.8	−.5	−.1	+3.7	0.0	−1.6	−2.6	−3.2	+20.5
Energy		+5									
5) Agriculture	2^0, −20%p	+42	−7.8	−1.4	−4.9	−4.8	−5.6	+3.5	−4.9	−7.0	+38.3
Durables		+10									
6) Agriculture	1^0, −10%p	+22	−6.8	−6.0	+2.0	+3.7	+.5	−.8	−4.8	+1.1	+23.9
Energy		+5									
Services		+5									
7) Agriculture	2^0, −20%p	+42	−12.8	−4.8	−.6	+5.0	−1.4	+5.4	−7.4	+1.1	+38.9
Energy		+9									
Durables		+10									
Services		+10									

TABLE 3—A COMPARATIVE ANALYSIS OF PARTIAL-EQUILIBRIUM WELFARE MEASURES

Scenario	Laspeyre's Price Index	*CV* Relative to Income: Low	High	Proportionate Error in Approximate Welfare Measures[a]: *PE*	*VPE*	*HPE*	ΣPE	$\overline{HPE}$
A. Single Sector								
1) Agriculture	1.001	.025	.014	.000	.487	−.564	–	–
2) Agriculture	1.003	.128	.072	.045	.627	−.657	–	–
3) Agriculture	1.006	.231	.125	.081	.688	−.691	–	–
B. Multiple Sector								
4) Agriculture Energy	1.006	.170	.113	−.325	.449	–	−.191	−.184
5) Agriculture Durables	1.015	.303	.284	−.411	.141	–	−.259	−.256
6) Agriculture Energy Services	1.012	.219	.197	−.466	.245	–	−.349	−.341
7) Agriculture Energy Durables Services	1.029	.421	.416	−.631	.004	–	−.393	−.386

[a]These partial-equilibrium welfare measures are distinguished according to single-market analyses (*PE*); vertically integrated analyses (*VPE*) involving the product market and relevant factor markets; or horizontally connected analyses involving the directly affected final goods markets ($\overline{HPE}$). ΣPE = the sum of the partial-equilibrium estimates.

puted to take place under each set of conditions in relationship to the baseline solution. Table 2 also reports the percentage change in each commodity's price, while Table 3 summarizes the overall price level impact with a Laspeyres price index. Table 3 includes the correctly calculated compensating variation relative to household incomes at the base case for the low- and high- income households.

II. Welfare Measures

Our analysis of partial-equilibrium measures of welfare change due to the specified effects of a climate change focuses on the price changes from the base solution, computed to arise with our CGE model. The welfare measures considered are defined as the Hicksian compensating variation (CV) for alternative price changes. Our partial-equilibrium CV measures are specified by varying the number of commodities whose price changes are recognized. As Table 2 indicates, each scenario (single and multiple sector) leads to changes in the vector of commodity prices. Assume, for example, the change is from $(P_1^0, P_2^0, \dots, P_9^0)$ to $(P_1^1, P_2^1, \dots, P_9^1)$. If we designate $E(\cdot)$ as the Hicksian expenditure function consistent with our utility function, then a single-sector, partial-equilibrium measure, CV_{PE}, would change only that sector's price, holding all others at the initial price levels as defined (for sector 1) in

$$(4) \quad CV_{PE} = E\left(P_1^0, P_2^0, \dots, P_9^0, U^0\right) - E\left(P_1^1, P_2^0, \dots, P_9^0, U^0\right).$$

We have considered three types of partial-equilibrium measures: 1) single-market (PE) analysis described above; 2) a multiple-market analysis with a vertically integrated version of extended partial-equilibrium analysis (VPE), where the prices of the specified consumption good and all of its inputs change to reflect the impacts of the climatic change; and 3) a multiple-market analysis with a horizontal connection of markets with the directly affected commodities prices changing for the welfare measurement. These last two measures follow from Richard Just et al.'s suggestions (1982). The last was formulated in two ways. For our scenarios that have direct changes in only one sector we specified it to involve all final consumption goods (HPE). When the scenarios imply simultaneous impacts on several sectors, it was defined to include only the directly affected final goods' prices ($\overline{HPE}$).

Before discussing the results we should acknowledge that while this approach does allow an evaluation of the importance of all the indirect effects for welfare measurement, it also places the partial-equilibrium welfare measures selected for evaluation in their "best light."[6] This is so because it assumes knowledge of *both* the correct expenditure function (for the welfare measures) and the correct price changes for all commodities that are considered in the analysis. In practice, as Whalley's 1975 study illustrated, we must estimate both the price change and a function to describe consumer preferences as a part of the partial-equilibrium modeling. Thus there are additional sources of error from each of these tasks. We do not consider them here. Rather we propose to gauge the importance of these indirect effects without the potential confounding errors introduced by other measurement tasks.

Table 3 reports the proportionate error found by comparing each partial-equilibrium measure of the welfare change (for each scenario) to the correct general equilibrium measure in the aggregate (i.e., summed across the two domestic households in each economy). That is, if CV_{PE}^k designates the compensating variation measured using a subset of the prices for scenario k, and CV_{GE}^k the correct general equilibrium measure, then our index of the error, EI_k, is given as

$$(5) \qquad EI_k = \left(CV_{GE}^k - CV_{PE}^k\right)/CV_{GE}^k.$$

The results for the single-sector scenarios indicate a clear preference for the partial-equilibrium, single-sector measure. Even with large unit cost impacts in that sector, the error in the partial-equilibrium measure represented a small understatement of the effect (i.e., less than 10 percent). By contrast, the extended partial-equilibrium measures have very large errors, even for *small* unit cost changes. However, we should emphasize that

[6] Of course, our proposed partial-equilibrium measures are just a sample of the possibilities. Each model will ultimately reflect the analyst's judgments in an attempt to best represent the important general equilibrium within a more restricted framework.

these findings are specific to the CGE model and scenarios we formulated.

Our overall work in this area (see our paper, 1985a) has considered several specifications for CGE models. In one case, with a model of comparable size and detail to the one reported here but for a less developed economy, the single-sector measures were uniformly worse (for comparable exogenous impacts) than the extended partial-equilibrium measures. Moreover, the magnitude of the errors were larger, often exceeding a 100 percent understatement of the welfare impacts with the single-sector *PE* measures. Given these discrepancies, it is essential to consider the features of the model's structure and of the scenarios that may have led to these results.

The first potential feature can be understood by referring to the results in Table 2. For the benchmark solution, agriculture accounts for about 14 percent of domestic output in this model. Nonetheless, the indirect effects of the climate impacts on other consumer commodities are calculated to be small in relationship to the primary effect on agriculture. Equally important, they are all the same sign and about the same magnitude. Thus, the set of consumer goods can be approximately treated as a Hicksian composite commodity. Under this assumption the use of a *PE* measure would largely represent a mistake in the magnitude specified for the change in price of the agriculture commodity *relative* to this composite good. If this interpretation is reasonable, we would expect welfare measurement errors to be related to the direction and magnitude of the error in the specified relative price change and it is. Using this framework, a single-sector *PE* approach would understate the increase in the relative price of agriculture goods. Thus, we would expect the *PE* welfare measure to understate the *CV* change. This is exactly what our results suggest.

In our analyses with the other CGE model referred to earlier, the economic sectors were more closely interconnected (with several produced commodities serving as intermediate inputs in multiple sectors). Moreover, in this case agriculture is more than twice as large a contributor to domestic output, accounting for about 30 percent. As a consequence, single-sector changes of comparable size in agriculture had larger indirect effects across the final consumption commodities. The sign and magnitude of the price impacts were not consistent across sectors. Consequently, the same Hicksian composite commodity assumptions would not be upheld in this case. Single-sector, partial-equilibrium measures would have been expected to be flawed, but the exceptionally large errors could not have been anticipated a priori.

This same framework also helps to understand our findings with the multiple-sector scenarios. In two cases (scenarios 4 and 6) the price effects (both the direct changes and approximate size of the indirect changes) are similar to those for one of the single-sector cases (scenario 2). The importance of the change as a fraction of the domestic households' incomes is also similar. Yet, the errors in the single-sector *PE* welfare measures range from 30 to nearly 50 percent overstatements. A part of the answer may well lie in the disparity in the directions of the indirect effects across sectors. Prices do not move in the same direction. In addition, while the average absolute magnitude of the percentage price change is about comparable to that experienced with the second scenario, the disparity across sectors is substantially greater.

Unfortunately, there does not appear to be a clear explanation for the performance of the extended partial-equilibrium measures. Presumably the answers to the performance of each measure lie in the patterns of relative price change in each scenario and importance of each commodity in the domestic households' expenditures in relationship to the definition of the commodities involved in the respective partial-equilibrium extensions.

On the second aspect of our application —its implications for economic impact analyses of a climate change—our findings should be interpreted only in qualitative terms because the model was not intended to authentically represent a specific economy. Nonetheless, they provide the first (to our knowledge) evidence of potentially large and

mixed price effects of a CO_2 induced climate change. Consequently, they (especially the multiple-sector scenarios) imply that focusing on selected sectors may give a very misleading description of the economic impacts of such a large scale change in the environmental conditions affecting production activities. In addition, they identify the possibility that the distributional implications of such changes may also be important; though clearly our results follow from the fact that climatic change affects the price of a commodity that is a large fraction of the low-income household's budget and the assumed patterns maintained for ownership of the CGE economy's resources across households.

III. Implications

Our results have confirmed a priori intuition with respect to the impacts of economy-wide shocks on the performance of single-sector, partial-equilibrium welfare measures. Since our explanations of the conditions leading to the cases of both the good and the poor performance are largely qualitative, one must consider the "value-added" from exercises such as this one. We believe there are at least two potential contributions of our approach. First, an important assumption in recent theoretical analyses of general equilibrium measures of welfare losses is proportionality in distortion changes. Diewert (1985), for example, found that if all distortion vectors increase proportionately then the approximate loss (measured with an Allais-Debreu index) increases quadratically. Our scenarios clearly indicate that proportionality in price movements in response to economywide impacts (or distortions) may in itself be a quite restrictive assumption.

Second, while qualitative judgments on the performance of partial-equilibrium welfare measures have clearly been a part of applied welfare economics for two decades, quantitative information on the exact magnitude of the errors introduced by ignoring indirect effects has been largely nonexistent. Our analysis suggests that the errors associated with these effects can be quite large for plausible scenarios and therefore it motivates further research to isolate the specific economic features that lead to these large effects.

Of course, it is also clear that the CGE framework provides a simple laboratory for evaluating a number of aspects of the practices of applied microeconomics, and thereby of adding a more precise quantitative basis for the judgments that are inherent in virtually all applied work.

REFERENCES

Allais, M., *A La Recherch d'une Discipline Economique*, Vol. I, Paris: Imprimerie Nationale, 1943.

Bach, Wilfred, Pankrath, Jurgen and Schneider, Stephen H., *Flood-Climate Interactions*, Boston: D. Reidel, 1981.

Boadway, Robin W., "Cost-Benefit Rules in General Equilibrium," *Review of Economic Studies*, July 1975, *42*, 361–74.

Debreu, G., "The Coefficient of Resource Utilization," *Econometrica*, July 1951, *19*, 273–92.

Diamond, P. A. and Mirrlees, J. A., (1971a) "Optimal Taxation and Public Production I: Production Efficiency," *American Economic Review*, March 1971, *61*, 8–27.

______ and ______, (1971b) "Optimal Taxation and Public Production II: Tax Rules," *American Economic Review*, June 1971, *61*, 261–78.

Diewert, W. Erwin, "Cost-Benefit Analysis and Project Evaluation: A Comparison of Alternative Approaches," *Journal of Public Economics*, December 1983, *22*, 265–302.

______, "The Measurement of Waste and Welfare in Applied General Equilibrium Models," paper presented at NBER Applied General Equilibrium Workshop, Stanford University, April 12–13, 1985.

Eastwood, D. and Craven, J., "Food Demand and Savings in a Complete Extended Linear Expenditure System," *American Journal of Agricultural Economics*, August 1981, *63*, 544–49.

Edlefsen, Lee E., "The Deadweight Loss Triangle as a Measure of General Equilibrium Welfare Toss: Harberger Reconsidered," unpublished paper, University of

Washington, 1983.

Firor, John W. and Portney, Paul R., "The Global Climate" in P. R. Portney, ed., *Current Issues in Natural Resource Policy*, Baltimore: Johns Hopkins University Press, 1982.

Harberger, Arnold C., "Professor Arrow on the Social Discount Rate" in G. G. Somers and W. D. Woods, eds., *Cost-Benefit Analysis of Manpower Policies*, Kingston: Industrial Relations Centre, Queens University, 1969, 81–88.

______, "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," *Journal of Economic Literature*, September 1971, *9*, 785–97.

Hare, F. Kenneth, "Climate Variability and Change" in R. W. Kates et al., eds., *Climate Impact Assessment*, SCOPE, Vol. 27, New York: Wiley & Sons, 1985.

Harrison, Glenn W. and Kimbell, Lawrence, "How Reliable is Numerical General Equilibrium Analysis?," unpublished manuscript, University of Western Ontario, January 1983.

Hazilla, Michael and Kopp, Raymond J., *Substitution Between Energy and Other Factors of Production: U.S. Industrial Experience, 1958–74*, Vol. 1, Final Report to the Electric Power Research Institute, Washington: Resources for the Future, August 1982.

Just, Richard E., Hueth, Darrell L. and Schmitz, Andrew, *Applied Welfare Economics and Public Policy*, Englewood Cliffs: Prentice-Hall, 1982.

Kokoski, Mary F., "A General Equilibrium Analysis of the Measurement of the Economic Impacts of Climatic Change," unpublished doctoral dissertation, University of North Carolina-Chapel Hill, 1984.

Kokoski, Mary F. and Smith, V. Kerry, (1985a) "A General Equilibrium Analysis of Partial Equilibrium Welfare Measures," discussion paper, Vanderbilt University, 1985.

______ **and** ______, (1985b) "General Equilibrium Welfare Measurement: A Cautionary Note," Working Paper No. 85-W-21, Vanderbilt University, April, 1985.

Lluch, C., Powell, A. and Williams, R., *Patterns in Household Demand and Saving*, New York: Oxford University Press, 1977.

McKenzie, G. W., *Measuring Economic Welfare: New Methods*, New York: Cambridge University Press, 1983.

Mansur, A. and Whalley, J., "Numerical Specification of Applied General Equilibrium Models: Estimation, Calibration and Data," in H. Scarf and J. Shoven, eds., *Applied General Equilibrium Analysis*, Cambridge: Cambridge University Press, 1984.

Scarf, Herbert, *The Computation of Economic Equilibria*, New Haven: Yale University Press, 1973.

Shoven, John and Whalley, John, "General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U.S.," *Journal of Public Economics*, Nos. 3/4, 1972, *1*, 281–321.

______ **and** ______, "Applied General Equilibrium Models of Taxation and International Trade," *Journal of Economic Literature*, September 1984, *22*, 1007–51.

Thompson, Louis M., "Weather Variability Climatic Change, and Grain Production," *Science*, May 1975, *188*, 535–41.

Waggoner, Paul E., "Agriculture and a Climate Changed by More Carbon Dioxide," in *Changing Climate: Report of the Carbon Dioxide Assessment Committee*, Washington: National Academy Press, 1983, ch. 6.

Whalley, John, "How Reliable is Partial Equilibrium?," *Review of Economics and Statistics*, August 1975, *57*, 299–310.

National Academy of Sciences, *Changing Climate: Report of the Carbon Dioxide Assessment Committee*, Washington: National Academy Press, 1983.

Comparative Productivity: The USSR, Eastern Europe, and the West

By ABRAM BERGSON*

This paper compiles comparative measures of output per worker in 1975 in four socialist and seven Western mixed-economy (WME) countries, and explores sources of observed differences between the two groups of countries in that regard. Such differences seem explicable only partially by reference to differences in per worker capital stock and farm land. A residual disparity of 25–34 percent in favor of WME countries appears to testify to superior efficiency in the latter.

In previous writings,[1] I compared in a more or less aggregative way output per worker in the USSR and several Western mixed-economy (WME) countries in 1960, and explored sources of observed differences in that aspect. Focusing on a more recent date, 1975, this essay attempts to extend this earlier work in order to embrace, in addition to the USSR, several Eastern European countries: Poland, Hungary, and Yugoslavia. I am also able to consider more WME countries than before, and to take account of further thoughts on methodology.

Among the possible sources of productivity differences between the groups of countries in question, one of particular interest is the difference in prevailing economic systems. The issue concerning the comparative efficiency of socialism that is posed is notably complex, but my previous studies may have limited speculation on it. Hopefully this one will serve in the same way.

An attempt to update and extend the reach of my previous studies seemed in order in view of the completion of another major phase (III) of the ongoing World Bank International Comparison Project (ICP) on relative national income. The latest work (Irving Kravis et al., 1982) provides systematic data on comparative real national income of the sort needed here for 1975, for numerous socialist as well as WME countries. The USSR is not among the socialist countries considered in ICP–III, but a careful inquiry into the relative national income of that country and the United States for a recent year, 1976, has become available elsewhere.

The ICP has compiled comparative data on national income not only for Poland, Hungary, and Yugoslavia, but for Romania. Because of limitations of available statistics on related matters of concern, I reluctantly omitted that country from this inquiry. Among WME countries covered by ICP–III, I limit myself to seven that are OECD members: the United States, Germany (the Federal Republic), France, Japan, the United Kingdom, Italy, and Spain. These countries do not vary as widely as might be desired regarding their development stage (the possible import of that matter will become evident), but inclusion of still other OECD members covered by ICP–III would have magnified an already onerous undertaking

*Harvard University, Cambridge, MA 02138. A separate Appendix, giving details on sources and methods for deriving basic data, is available to readers upon request. I am indebted to Zvi Griliches and anonymous readers for helpful comments on earlier versions, and have also benefited from discussion with Michael Kerens and Gur Ofer. In utilizing national statistics for different countries, I have been aided in more than one way by numerous persons and agencies. Particular thanks are due to Edmond Malinvaud of INSEE, Paris; the Statistisches Bundesampt, Wiesbaden; Takao Fukuchi of the Economic Planning Agency, Tokyo; Thad P. Alton, of L. W. International Financial Research, New York; Ivo Vinski and Michael L. Boyd. Mary Towle has provided her usual sterling secretarial assistance. Daniel Quah and Marc Rebick also served by carrying out regression computations.

[1] Chiefly 1971 and 1972 studies, reprinted in somewhat revised form in my book (1978). See also my studies (1964; 1968), and, on methodological aspects, my paper (1975).

without much substantive gain. I also pass by a number of Third World countries that the ICP covers.

Since the completion of ICP–III, results of other related calculations, chiefly in connection with the still ongoing Phase IV of that program, have already begun to appear. For various reasons, comparison with the results of phase III is difficult, but may shed some light on the reliability of the latter.

A socialist economic system is understood here in the conventional, though admittedly controversial, way as one where ownership of the means of production is predominantly public. Among the countries in question, though, economic working arrangements (i.e., institutions, policies, and procedures determining resource use) often differ from one country to another. Thus, despite numerous highly publicized reforms, planning in the USSR and Poland in the mid-1970's was still broadly of the centralist sort, stressing bureaucratic as distinct from market coordination of production units, that originated in the USSR under Stalin. In Hungary that form of planning has given way, since 1968, to one stressing market processes, though probably not to the degree often supposed. Yugoslavia's economic system seems almost continually in flux, but here, too, market processes have tended to be in the ascendant. Also, under so-called industrial self-management, workers, at least in principle, are ultimately in control of the production unit (for which reason, Yugoslavs prefer to refer to their system as one where there is "social" rather than "public" ownership of the means of production). In both Poland and Yugoslavia, agriculture is still predominantly of the peasant as distinct from collectivized sort. Such facts should be borne in mind. Of course, economic working arrangements also often differ markedly among WME countries, and that, too, must be considered.

Also to be noted is the fact that in 1975 a poor harvest probably cost the USSR a few percent of its GNP. The year 1975 also tended to be a recession one in the West. Among countries considered, output and employment responded variously. Three countries, the United States, Germany, and Italy, suffered absolute declines in GDP of 0.9, 1.8, and 3.6 percent, respectively.

I. Methodology

Resistant as comparative efficiency is to measurement, it is still advisable to be clear that I focus, as before, on production efficiency as manifest in realization of production possibilities. In the case of labor, however, attention is directed only to the use made of employed endowments, and my findings will have to be read accordingly.

In seeking insight into intersystem differences in efficiency, as so understood, I focus on this equation:

$$\text{(1)} \quad \log y = A \log k + B \log l + Md + Q.$$

Here, A, B, M and Q are constants; y is output per worker, relatively to that in the United States; k and l are capital and land per worker, similarly calculated; and d is a dummy variable denoting socialism or its absence. The constants in (1) are to be evaluated by regressing y on k, l, and d, in the light of available observations on their magnitudes. Of the different constants, M, the coefficient of d, is necessarily of particular interest. I interpret the regression relation that is obtained, however, and hence also M, in a rather novel way.

In applying (1), I evidently assume that production both in the WME countries and under socialism conforms to a log-linear version of the Cobb-Douglas formula. According to the usual understanding, that would mean determining the contribution (positive or negative) of socialism to y after normalizing, in conformity with that formula, for differences in factor inputs, as represented by k and l. The constants A and B are seen correspondingly as having the familiar status of output elasticities or comparative "earning shares" imputable to capital and land, respectively. And that is also the understanding here, but in the present context k is viewed not only as the input of capital per worker but as an indicator—it is an appealing one, I think—of a country's stage of development. In normalizing for k, therefore, I allow not only for differences in per

worker capital as such, but also for additional forces affecting output per worker that are associated with the development stage.

Such forces must sometimes affect output per worker rather than efficiency; for example where, as commonly must be so, the technological knowledge of a less advanced country lags behind that of a more advanced one because of its tendency to borrow rather than generate new knowledge. Such a lag has an interest of its own, but would not exemplify efficiency in the usual sense envisaged here. Sometimes, though, efficiency too must be affected, as where a historically distorted resource allocation can be remedied only as the development process unfolds. But for our purposes, normalization is appropriate in either case, and that seemingly is accomplished by inclusion of k as an independent variable. With that, should M differ significantly from zero, that should testify that, depending on its sign, socialist countries tend to be more or less efficient than WME ones at the same stage.

Development stage is open to more than one construction, and capital per worker is not the only criterion that might be used to gauge it. That measure has a distinct advantage here, however, over a rival often employed, agriculture's share of the labor force. Thus, it does not itself depend on efficiency, as the agricultural share in the labor force clearly does. Other familiar criteria, such as output per worker and output per capita, evidently cannot serve us at all.

Formula (1) presupposes that each of the three inputs embraced is homogeneous, but in principle is easily elaborated to allow for heterogeneity. As usually so in computations such as in question, however, we must settle for something less than the ideal kind of data required, which for each input, including labor, entails the same sort of earnings-share-weighted logarithmic aggregation of varieties as that which is in effect applied in (1) in totalling the inputs of capital and land. Thus, labor will be calculated simply in physical units, though an attempt is made to allow in a conventional way for quality. Land here reduces to that employed in agriculture, and that, too, will be computed in physical units. I try, however, to allow for quality in one outstanding case, the USSR.[2] I refer, as usually done, to a linear sum of the price-weighted volumes of different capital items.

For capital, the volumes in question properly are not of stocks but of services, and correspondingly reference should be not to the prices of capital goods but to rental rates. In practical work, however, one is usually offered a choice between two procedures, neither one of which is entirely satisfactory: to take services as proportional to gross assets, on the one hand, or to net assets, on the other. Partly because of the greater availability of relevant data, I opt for the former course, but we cannot ignore altogether that services tend to decline with age, even if not always commensurably with depreciated value. The rate of utilization also matters, though here, too, quantification is difficult.[3]

[2] Farm land may perhaps be viewed here as representing not only itself but other nonreproducible capital, which is not otherwise accounted for. As between the socialist and WME countries studied, in any event, there is little basis to suppose that such capital is relatively much more abundant in one case than in the other.

[3] Intercountry differences in utilization of the capital stock are manifest in diverse ways, but principally in differing shares of capacity that may be completely inoperative, and, for capacity that is in any degree operative, in differing labor hours and use of shift work. Productivity is calculated in this essay by reference to employed workers only. In the case of the capital stock, arguably a parallel treatment calls for adjustment to allow for disparities in services on all of the indicated counts. While relevant to efficiency, the import of the differences in question from that standpoint would then be reserved for separate inquiry.

Logical as it is, such procedure can be only very partially applied here. But a principal limitation is probably our inability to allow, except in respect of the resultant shortening of hours, for the reduction in capital services employed in WME countries that was associated with the 1975 recession. Perhaps it is as well, though, to have WME productivity levels deflated in that way by inclusion of recession-induced underutilization of capital.

However that may be, a question remains as to how capital services vary with indicators of utilization such as in question. If only for illustrative purposes, I shall take the relation to be one of proportionality. While often done, that is likely to overstate the variation in

I have construed the Cobb-Douglas formula in the usual way, excluding economies of scale. So far as scale economies are realized, they are reflected in output per worker, and what remains to be accounted for here is differential opportunities for such economies. Development stage apart, such opportunities presumably depend in a degree on country size. In applying (1), I accordingly explore a variant where an additional parameter, population, is introduced. The regression computation also allows another useful modification: rather than assigning but one dummy variable for all socialist countries together, such a variable is assigned separately to each.

With available data, each country's output can be compared with that of the United States in terms of both its own and U.S. prices (similarly for each country's capital stock). While prices nowhere conform fully to well-known theoretic desiderata for computations such as ours (compare my paper, 1975), they often diverge egregiously from such desiderata under socialism. It thus seems best to focus here primarily on calculations where valuations are in U.S. prices, but it will still be of interest to see how the results are affected when valuation is in other country prices.[4]

With only 11 observations, opportunities to explore alternatives to the Cobb-Douglas production function are limited. I nevertheless experimented with a partial shift to a CES formula, specifically one entailing retention of Cobb-Douglas to aggregate capital and land, but leaving the elasticity of substitution (σ) unconstrained otherwise. The resulting σ turns out, however, to be oddly negative, and not surprisingly fails to be significantly different from unity at the 10 percent level; or at the 20 percent level when labor is adjusted for quality. Reassuringly, though, results in respect of M are fully in accord with those from application of (1), to be presented below.[5]

I consider performance in the economy generally, but for familiar reasons exclude the services of personnel engaged in the provision of education and health care, and delimit correspondingly inputs of capital. In the WME, labor and capital employed in government administration, including defense, and the resultant output, are also

services. As easily seen, that follows from the fact that the appropriate rental rate for any given asset is given ideally by the sum of interest and depreciation on it, while any capital gain, if of an expected sort, is to be deducted (compare Laurits Christensen et al., 1981). The volume of services rendered by any given gross stock, it must also be considered, might be affected by differences in the physical structure of the capital stock, and resulting differences between relative prices and gross rentals.

Formula (1) also presupposes that output in each country is a single homogeneous product. So far as there are many products, theory allows for more than one kind of aggregation (my paper, 1975), but I must in any event opt for the usual price-weighted linear form.

[4]I am thus in effect exploring implications of both Laspeyres and Paasche index numbers—to give them their proper names. That is a somewhat cumbersome procedure, but, given the highly dubious nature of the socialist prices, it seems preferable to the alternative of focusing simply on one or another summary composite such as Fisher's "ideal" index or the relatively novel multilateral "superlative" index. The latter measures are nevertheless still of interest but results of use of the ideal index must fairly clearly be more or less congruent with those indicated by our Laspeyres and Paasche indexes. As for the superlative index, at least at levels now of interest, resultant measures of output appear to conform quite well to those obtained by use of the ideal formula: see, for example, Christensen et al. (p. 89) and further ideal indexes in the source of ICP–II data cited there.

[5]I applied this formula:

$$\text{(a)}\quad \log y = (1/\alpha)\log[C+(1-C)\bar{k}^{\alpha}] + Md + Q.$$

Here $\bar{k}$ is the indicated composite of reproducible capital and land. In aggregating the two inputs, I take the needed output elasticities to be proportional to those obtained by application of (1). The terms y, d, M, and Q are understood as in (1), while α is a parameter, such that

$$\text{(b)}\qquad \sigma = 1/(1-\alpha).$$

These are the resulting magnitudes of α and, for later reference, those for M, with associated T values indicated parenthetically: with labor not adjusted for quality, α, 2.425 (1.73) and M, -0.363 (8.81); with labor so adjusted, α, 2.448 (1.13); M, -0.300 (6.09). In inferring significance levels, I reduce degrees of freedom by one in order to allow for the use of previously derived factor coefficients in calculating the capital-land composite.

TABLE 1—GROSS DOMESTIC MATERIAL PRODUCT (GDMP), GROSS REPRODUCIBLE CAPITAL STOCK (GRCS), AND FARM LAND, PER WORKER, SPECIFIED COUNTRIES, 1975[a]
(USA = 100.0)

	Per Worker, Adjusted for Nonfarm Hours			Per Worker, Adjusted for Labor Quality		
Country	GDMP	GRCS	Farm Land	GDMP	GRCS	Farm Land
USA	100.0	100.0	100.0	100.0	100.0	100.0
FRG	94.1	111.0	15.4	108.3	127.8	17.7
	(76.9)	(89.0)		(88.5)	(102.4)	
France	90.7	81.7	39.4	104.4	94.0	45.4
	(71.6)	(69.3)		(82.4)	(79.7)	
Italy	71.0	71.4	28.8	81.6	82.0	33.1
	(56.7)	(68.2)		(65.2)	(78.3)	
UK	68.6	78.7	14.8	75.0	86.1	16.2
	(57.0)	(68.5)		(62.4)	(75.0)	
Japan	56.9	65.4	3.7	64.1	73.8	4.2
	(41.4)	(50.3)		(46.7)	(56.7)	
Spain	56.0	41.3	58.1	67.8	50.0	70.3
	(40.7)	(31.9)		(49.3)	(38.7)	
USSR	47.4	57.8	81.8	57.6	70.3	49.7[b]
	(30.8)	(45.2)		(37.5)	(54.9)	
Hungary	43.4	50.4	42.6	52.8	61.3	51.7
	(31.6)	(41.2)		(38.4)	(50.1)	
Poland	36.2	34.1	33.2	44.7	42.1	41.1
	(26.2)	(28.8)		(32.4)	(35.5)	
Yugoslavia	33.9	29.3	37.3	42.0	36.3	46.2
	(26.4)	(26.7)		(32.7)	(33.1)	

[a]For GDMP and GRCS, each comparison with USA is in U.S. prices; the comparisons in other countries' prices are shown in parentheses. For concepts more generally, and sources and methods, see text and the Appendix (available upon request).
[b]Farm land as well as labor adjusted for quality; see text.

properly omitted. By reference to the funding involved, particularly whether financing is through the so-called "government" budget, one can delineate and exclude labor and capital employed and the associated output in an essentially similar sphere in the socialist economies studied. So as to allow, if only arbitrarily, for the productive use of highways and streets, I omit only one-half the capital of those funds, although that is in the WME almost entirely and in our socialist economies entirely, of a governmental kind.[6]

Because of the special nature of housing services, I also omit them from output, and exclude from inputs the capital involved in their provision. No attempt is made, however, to omit from labor inputs the relatively limited number of workers so employed. I understand by "selected services" the foregoing diverse omissions. The economy generally, exclusive of those aspects, is referred to as "material sectors," though services apart from selected ones are, of course, included.

II. The Data

I have assembled here (Table 1) the basic data on output and inputs in 1975 to which (1) is to be applied. For all countries, I refer to output in material sectors; hereafter gross domestic material product (GDMP). The gross reproducible capital stock (GRCS) is of corresponding scope, and so too is the level of employment to which output, the

[6]More generally, while excluding government administration from the scope of my calculations, I try, data permitting, to include in their entirety, and without regard to their ownership or administrative status, all transportation infrastructures other than highways and streets, and also the stocks employed in irrigation and conservation; and urban water supply. Urban sewage works, however, are omitted. On the scope of government administration, compare Kravis et al. (pp. 66–68).

TABLE 2—GROSS DOMESTIC PRODUCT (GDP), GROSS DOMESTIC MATERIAL PRODUCT (GDMP), EMPLOYMENT AND GROSS REPRODUCIBLE CAPITAL STOCK (GRCS), AND FARM LAND, PER CAPITA, SPECIFIED COUNTRIES, 1975[a]
(USA = 100.0)

			Employment per Capita					
Country	GDP per Capita	GDMP per Capita	All Sectors	Material Sectors	Material Sectors Adjusted for NFH	Material Sectors, Adjusted also for Labor Quality	GRCS per Capita	Farm Land per Capita
USA	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
FRG	88.3	90.9	100.6	101.4	96.6	84.0	107.3	14.8
France	89.5	92.2	99.1	100.2	101.7	88.3	83.0	40.1
Italy	60.7	61.3	87.4	87.9	86.4	75.2	61.6	24.9
UK	73.5	67.2	109.4	102.1	98.0	89.6	77.2	14.5
Japan	75.5	82.8	115.0	128.6	145.5	129.0	95.2	5.4
Spain	62.0	64.6	91.3	101.0	115.4	95.4	47.7	67.0
USSR	60.2	60.0	125.2	132.0	126.5	104.1	73.2	103.5
Hungary	56.3	61.1	120.1	135.1	140.6	115.6	70.9	59.8
Poland	54.2	54.8	126.7	143.2	151.5	122.7	51.6	50.4
Yugoslavia	41.2	41.5	107.3	121.0	122.4	98.8	35.9	45.6

[a]Output and capital stock in U.S. prices. For concepts more generally and sources and methods, see text and the Appendix.

capital stock and farm land are related. That is so whether employment is adjusted for differences in nonfarm hours (NFH) or further adjusted for differences in labor quality.

Details on sources and methods used in compiling those data are set forth in an Appendix (available upon request); a summary account of their provenance, however, will serve among other things to bring out limitations that must be considered in applying (1). Sources and methods are much the same whether GDMP and GRCS are valued in U.S. or other country prices, but it facilitates discussion if I focus on data where the former valuation is used, and where related elements underlying Table 1 are as shown (Table 2).

Output. For all countries, GDMP is calculated by exclusion of services from GDP. For the GDP, for all countries other than the USSR, as indicated, I draw on ICP–III. The GDP of the USSR, relatively to that of the United States, is derived essentially from Imogene Edwards et al. (1979). The ICP calculations have been widely greeted as of outstanding merit, and those of Edwards et al. are also the result of an unusually systematic inquiry. In comparing Soviet and U.S. output in terms of U.S. prices, though, Edwards et al. had to grapple somehow with the notable volume of defective or otherwise substandard goods that are produced in the USSR, apparently often to be sold at prices of standard products. Careful as the computations were, it seems doubtful that they could have allowed sufficiently for that feature. More generally, because of residual qualitative deficiencies in Soviet products that couldn't be accounted for, as the authors themselves acknowledge, their calculations could significantly overstate Soviet output.[7]

The production of substandard goods in Eastern Europe has yet to be explored in any depth. Such production on some scale has apparently been a feature wherever centralist planning prevails, and that scheme still operates in Poland. Performance regarding prod-

[7]In addition to Edwards et al., on the question of product quality as it affects calculation there of the comparative volume of Soviet output, see the sources cited in the Appendix on the ruble-dollar ratios compiled for investment goods, and Gertrude Schroeder and Edwards (1981) on those for consumer goods. On the consumption component of GDP, I have revised Edwards et al. to conform to Schroeder and Edwards.

TABLE 3—COMPARATIVE GROSS DOMESTIC PRODUCT AND EXPENDITURE PER CAPITA, ALTERNATIVE COMPUTATIONS, SELECTED COUNTRIES, 1975[a]
(AUSTRIA = 100.0)

	Prices: ICP–III, 1975				Prices: APC, 1975		ICP–IV, 1980
Country	U.S.	International	Austrian	Other Country's	Austrian	Other Country's	Extrapolated to 1975
GDP, less Net Exports, per Capita							
Austria			100.0	100.0	100	100	
Poland			80.9	64.9	67	54	
GDP per Capita							
Austria	100.0	100.0	100.0	100.0			100.0
Hungary	73.6	71.3	79.0	64.1			58.6
Poland	70.8	72.0	77.3	60.9			60.1
Yugoslavia	53.9	51.9	55.7	49.1			45.5
FRG	115.4	119.2					116.0 (119.5)
France	117.0	117.7					115.2 (117.6)
Italy	79.3	77.3					89.6 (83.5)
UK	96.1	91.9					103.7 (100.7)
Spain	81.0	80.3					81.7

[a]*Sources*: For ICP–III, Kravis et al.; for APC, A. Franz et al. The ICP–IV data for 1980 are from UN Statistical Commission and Economic Commission for Europe (1985), retrapolated to 1975 by reference, for WME countries, to GDP and population in OECD (1982a); for Poland and Hungary, to GNP and population in Thad Alton et al. (1983) and SEV (1981), and for Yugoslavia, to GNP in OECD (1982b) and to de facto population in Appendix Table 1 and as inferred from OECD (1982b). On ICP–III international prices, ICP–IV prices, and alternative figures in parentheses, see fn. 8.

uct quality in Hungary and Yugoslavia, though, is often assumed to have benefited from the increased reliance there on market processes.

We must also consider, however, further calculations of comparative output in Austria and Poland (APC) made collaboratively by the national statistical offices of the two countries; and of comparative output in Austria, Poland, Hungary, and Yugoslavia, made as part of ICP–IV. The latter calculations likewise involved a joint effort on the part of the countries concerned, though with Austria exercising primary responsibility. Because of the different valuation standards applied, and, in the case of ICP–IV, the focus on 1980, the import of these further inquiries for the reliability of our ICP–III data on comparative GDP in East European countries and the United States in 1975 in U.S. prices is not easy to judge. Juxtaposition of the alternative measures with related ICP–III results, however, seems to underline for Hungary and Yugoslavia as well as Poland, that the ICP–III data on GDP per capita are more likely to err in the direction of over- than of underestimation (Table 3). Further ICP–IV measures for Western European countries seem more or less consistent with those of ICP–III. For Italy and the United Kingdom, though, ICP–III may have underestimated GDP per capita somewhat in relation to other European countries.[8]

[8]In ICP–IV, comparative Austrian-Eastern European output is calculated essentially in terms of a weighted average of relative prices in the three Eastern European countries, Austria, and Finland. ICP–IV comparative data for Austria and Western European countries are calculated essentially in terms of average relative prices of ten EEC member countries. ICP–III international prices purportedly represent similar averages of global scope. While ICP–IV data for 1980 are retrapolated to 1975 by reference to comparative trends in real GDP (GNP) per capita, I also show parenthetically corresponding measures obtained after partial adjustment for

Employment. I utilize here data on numbers employed and NFH drawn mainly from OECD, other Western and socialist official sources, and ICP–III. The results are unavoidably inexact, but there is no basis to suppose a bias here might offset that very possibly affecting socialist GDP per capita. I allow in a usual, Denison-like way for labor quality, as that is affected by differences in hours, sex structure, and educational attainment, and also draw on Edward Denison (1979) for earnings weights needed in the computations. As is well-known, such calculations even at best have their limitations; because of deficiencies in available data, they are not at their best here.[9] For that reason, I apply (1) to data on output and inputs both before and after adjustment for labor quality.

Capital Stock. In calculating GRCS in terms of U.S. prices of 1975, I apply to data on stocks in national currencies purchasing power parities (*PPP*s) relating to producer's durables, construction and the GDMP generally that were used in the calculation of output in U.S. prices of 1975. The data on socialist output in terms of U.S. prices of 1975, though, may well be overestimates. If they are, the cause should be the overvaluation of socialist currencies in terms of dollars. Such an overvaluation, then, may also occur at this point, and correspondingly an overstatement of socialist stocks of reproducible capital per worker.

As for the gross reproducible capital stocks in national currencies, the dominant element here is fixed capital. For all Western countries except Spain, I could draw on results of substantial inquiries, usually of the perpetual inventory sort. While that procedure is subject to familiar limitations, resultant errors need not be in one direction rather than the other.[10] For Italy, gross stocks had to be estimated from data computed on net stocks. For Japan, I extrapolate from results of an official wealth survey for December 31, 1970. No data are at hand on Spain's fixed capital stock. Relatively to that of the United States, I take Spain's stock to be of the same magnitude as its aggregate 1960–74 volume of investment in fixed capital (excluding housing) is relatively to the corresponding U.S. total. Fixed capital investment (excluding housing) has lately been growing rapidly in Spain—over the decadal interval 1960–64 to 1970–74, by 10.6 percent yearly. In taking comparative investment volume during 1960–74 as a yardstick, I may have overestimated Spain's terminal stock.

For socialist countries, I draw on official data that derive in each case primarily from *ad hoc* surveys of fixed assets on hand at particular dates. The statistical agencies concerned apparently expand the survey results into serial data by reference to current new investment and retirements. Asset surveys such as are in question have, in the case of the USSR, been subject to close Western scrutiny with results that are reassuring as to their substantial reliability. More generally, the socialist official data have their limitations, but there should not be any systematically consequential overstatement.[11]

disparate volume trends in final use categories: see Eurostat (1983, p. 115); as confirmed by Hugo Krijnse Locker (letter of April 25, 1984), relevant figures cited there for Germany and Italy should be corrected to read (−) 2.9 and (+) 7.3, respectively.

Compared to ICP–III, both APC and ICP–IV have the virtue of using more voluminous price data in deriving *PPP*s needed to translate outlays from one currency to another. In ICP–III, though, a novel "country product dummy" (CPD) method was used to augment price quotations.

[9]Among other things, for Spain I assume average educational attainment of male workers is the same as for Italy. While the adjustment for labor quality, where GDMP and GRCS are in U.S. prices, has its limitations, the earnings weights at least properly reflect U.S. values. Where GDMP and GRCS are in other country prices, corresponding earnings weights are not at hand, so I apply again those reflecting U.S. values. One might wish to derive weights for different qualitative features, along with constants such as in (1), by regression analysis, but that is not feasible here.

[10]Among the more troubling features is the introduction, either implicitly or explicitly, of assumptions as to asset longevity that are often of doubtful validity. Simulations made for the United States and United Kingdom, however, suggest that resulting stocks are probably not as sensitive as might be supposed to error at this point: see U.S. Department of Commerce Bureau of Economic Analysis (1982, pp. T–13ff, 1, 197); J. Hibbert et al. (1977, pp. 124–25).

[11]On Soviet surveys of the fixed capital stock, see Raymond Powell (1979) and the sources cited there.

(Continued)

Although capital services cannot be expected to correspond to the depreciated value of an asset, as noted, they tend to decline to some extent with its age. Hence, relatively to the gross values that I record, services can be expected to vary so far as the assets in question differ in age. Such variation could be material here, since service lives of capital goods appear to be relatively lengthy in socialist countries. The resultant disproportion between services and gross capital stock, though, should not be too great, for at relevant tempos of growth the share of superannuated goods in the capital stocks of the socialist countries must be rather limited. There is more than that to the question of comparative services and gross capital stocks, but further biases that can be discerned in my measures of capital inputs at this point are by no means unidirectional. Overall, however, some relative overestimation of socialist capital services is not precluded.[12]

Farm Land. Understood essentially as arable land, including land under permanent cultivation, this is represented simply by its acreage as determined from standard sources. In calculating factor productivity with labor adjusted for quality, however, I also reduce farm land in the USSR by one-half, that discount being suggested by reference to U.S. climatic analogues (see my study, 1964, p. 379).

Comment. Relatively to each other, output and employment per capita have been found sometimes to vary rather differently for the whole economy and material sectors (Table 2). Such incongruities must be read, however, in the light of the relatively high U.S. prices of selective services employment, which in a number of countries tends markedly to inflate their product relatively to that of material sector workers (compare Kravis et al., pp. 191 ff.). The notably high socialist ratios of employment to population that are observed are, of course, mainly an expression of the well-known tendency in socialist countries to high rates of labor par-

Granting that these inquiries inspire confidence, the official data drawn on here for 1975 had to be derived from a survey for 1971–72. The official measures of "real" investment volume that must have been used in such a calculation have lately been held likely to overstate growth in that aspect. Very possibly, they do in a degree (see my forthcoming study), but for the 3–4 year interval for which their use is in question here, the effect on the official stock data for 1975 must have been quite limited.

The last fixed capital stock survey date before 1975 was for Hungary, 1968; Poland, 1960; and Yugoslavia, 1953. Should there be any upward bias in official data for real investment volume for these countries, it might seem the cumulative effect on stocks reported for 1975 could be significant. In official serial data, however, while stocks are initially valued at prices prevailing around the date of a survey, they subsequently are adjusted for price level changes. The adjustment seems clearly to have involved application of price index numbers that are complementary to the official measures of "real" fixed investment: compare with due allowance for presumable aggregation differences, the implied deflators for capital formation and stocks in UN national accounts yearbooks and successive editions of the statistical yearbooks on which I draw for capital stock data in the Appendix. In finally translating the official stock data, as necessary, to 1975 prices, I myself essentially apply implied investment deflators.

Should official measures of real investment volume be subject to an upward bias, therefore, the price index numbers used to update capital stock values must be subject to a downward one. On balance, as readily seen, the result should be an understatement of the fixed capital stock in 1975 prices. I have, nevertheless, been informed by Leszek Zienkowski of the Polish statistical office (letter of October 28, 1982) that the official data overstate Polish stocks. The grounds for that opinion are not clear. Possibly the Polish capital stock is relatively aged (see below). For Yugoslavia, the correspondence of capital stock inflation and investment deflation could be established only in respect of a capital stock revaluation from 1966 to 1972 prices, but judging from calculations of Ivo Vinski (see the Appendix), the fixed capital stock of 1975 could be some 8.8 percent greater than has been estimated here. While I rely primarily on official data, I have referred to Vinski to fill in some gaps.

[12]Differences in the relation of services to gross stocks also result from differences in the use of shift work and in working hours, and the inclusion in gross stocks, as calculated, of varying amounts of unfinished construction. From often notional calculations, I judge that with allowance for those aspects our measures of GRCS per worker, adjusted for NFH, might have to be further adjusted by these amounts, in percentage points with the United States as 100.0: FRG (−) 6.7; France (−) 0.8; Italy (−) 2.9; U.K. (−) 4.2; Japan 3.5; Spain 4.0; USSR (−) 0.6; Hungary 3.6; Poland 2.8; Yugoslavia 0.8. On the foregoing, and on services more generally, including the possible import of differences in service lives and growth, see fn. 3 and the Appendix.

ticipation.[13] Employment per capita is also high in Japan, however, and the relative labor input in that country is the greater because of the long hours there. Hours are also long in Spain, but overall hours are not too different in WME and socialist countries. Owing chiefly to their comparatively limited educational attainment, all countries other than the United States—both socialist and WME—experience a decline in per capita labor inputs compared to the USA when an adjustment is made for quality (Table 2). Here too, though, Japan is rather special since educational attainment there compares favorably with that everywhere but in the United States.

Much has been written lately about the so-called "second economy" of the socialist countries. As not always made clear, many of the undertakings referred to are quite legal, but by all accounts there has been a proliferation in the course of time of diverse quasi-legal (i.e., not officially approved or disapproved) and illegal economic activities. The under coverage of economic data that almost inevitably results, however, apparently affects inputs as well as outputs. Also, activities in question relate solely to private enterprise, while reported statistics are inclusive of the socialized sector which is of primary concern here. Of course, underreporting is not confined to socialist countries, but for the WME, too, the reported data retain interest.[14]

III. Results

With GDMP valued in U.S. prices, output per worker is found to vary widely among both WME and socialist countries, but for all the latter it falls distinctly short of levels attained in the former. With employment adjusted for NFH, output per worker among WME is at the least 56.0 percent of that of the United States, while among socialist countries it varies from 33.9 to 47.4 percent of the United States (Table 1). With employment adjusted also for quality, the corresponding minimal WME level is 67.8 percent, while among socialist countries output per worker ranges from 42.0 to 57.6 percent of the U.S. level. Chiefly because of the possible overvaluation of socialist currencies in terms of dollars, my calculation may understate the margin between WME and socialist countries in respect of output per worker.

Among diverse regressions computed on sources of that gap, I refer first to a number where employment has been adjusted for NFH but not for quality. Of these, priority is accorded one (Table 4, I–1) simply applying formula (1) to relevant data (Table 1).

[13] In this essay employment is understood to be of the domestic sort in the country concerned, and thus inclusive of foreign workers employed there. That is as it should be where employment is juxtaposed with "domestic," as distinct from "national," output. The corollary, though, is that employment for a country excludes its workers abroad. That should be considered in construing the cited data on employment per capita, especially those for Italy, Spain and Yugoslavia. In the case of the latter, workers abroad are particularly numerous.

Prescribed Western national income accounting procedures for dealing with earnings of foreign workers (UN, 1968, p. 93) are rather complex, and can hardly be adhered to always, even among Western countries. Available employment data also seem unlikely to be entirely congruent with GDP as actually computed. Among socialist countries, however, workers abroad are consequential only for Yugoslavia, and such workers are known to be generally excluded from Yugoslav employment data as recorded in this essay. Any error to speak of here thus should be on the side of improper inclusion of remittances in, and hence of overstatement of, Yugoslav GDP, and presumably of GDMP per worker.

[14] At least for the USSR, the undercoverage due to second-economy activities probably is not as great as often imagined. Taking retrospective emigre budgets for 1972–74 as a point of departure, Gur Ofer and Aaron Vinokur (1980, p. 51) conclude that omissions from urban household incomes may come to 3–4 percent of the GNP. The cited figures would not reflect significant illegal rural activities (probably mainly distilling), but second-economy output must often be overpriced. For Hungary there are indications that at the time studied, activities such as in question may have been relatively to GNP much more extensive than in the USSR. While our data should be inclusive in respect of the socialized sector, they must understate productivity there to some extent so far as socialist sector inputs are reportedly sometimes appropriated for private use, for example, state materials are pilfered for use in private construction.

TABLE 4—ALTERNATIVE REGRESSIONS WITH OUTPUT PER WORKER, ADJUSTED FOR NFH, AS DEPENDENT VARIABLE, 1975[a]

Regression	Scope of Dummy Variable	R^2	Regression Coefficient A	B	M	N	Q
I–1	Socialist Countries	.970	.560 (8.22)	.092 (3.85)	−.351 (5.77)		−.001 (0.018)
I–2	Socialist Countries	.976	.560 (8.22)	.092 (3.85)	−.397 (6.01)		.001 (0.018)
I–3	Socialist Countries	.976			−.331 (4.58)		
I–4a	USSR	.972	.644 (8.08)	.101 (4.27)	−.409 (5.38)		.036 (0.76)
I–4b	Hungary				−.343 (4.54)		
I–4c	Poland				−.247 (2.67)		
I–4d	Yugoslavia				−.227 (2.24)		
I–5	Socialist Countries	.969	.589 (7.60)	.096 (3.86)	−.351 (5.64)	−.023 (0.85)	−.013 (0.28)
I–6	Socialist Countries	.949	.645 (6.27)	.102 (2.63)	−.410 (4.20)		−.075 (1.01)

[a]The T-statistics are shown in parentheses, each having the sign of the coefficient to which it relates. R has been adjusted. For purposes of the computations, output and inputs per worker and population are expressed in relation to USA = 1.000.

The striking result is that with R^2 as high as .970, a dummy variable standing for socialism has a negative coefficient, M, significant at the 1.0 percent level. Judging from the magnitude of the coefficient, −.351, output per worker under socialism, tends to be 29.6 percent below that in a WME country.

The two other variables considered, capital and farm land per worker, also have the regression coefficients, A and B, respectively, that are significant at the 1.0 percent level. The coefficient for capital, .56, is distinctly larger than the corresponding "earnings share" that is often imputed to capital in factor-productivity studies for the more advanced WME countries. That is not surprising, since capital per worker represents here not only the corresponding input but also the stage of development. An increase in capital per worker thus predictably has a more pronounced impact on output per worker than factor earnings imputable to capital in an advanced WME country might indicate.

In I–2 and I–3 (Table 4) all is as in I–1, except that I now allow illustratively for the possible overvaluation of socialist national currencies in U.S. dollars. I consider here that as a result socialist capital stocks as well as outputs may be overstated. In I–2, I reduce both the output and capital per worker of socialist countries by 10 percent. In I–3, their output per worker is cut by 10 percent and their capital per worker by 20 percent. With either adjustment, all is essentially as in I–1, with the dummy variable for socialism still having a negative coefficient significant at the 1.0 percent level. The coefficient, however, is now a somewhat larger negative in I–2 and a somewhat smaller negative in I–3 than it was in I–1. Consequently, while socialism in I–1 underperforms a WME country by 29.6 percent, the corresponding shortfall under I–2 is 32.8 and under I–3, 28.2 percent. Not shown in the table is another computation which, while allowing for possible overvaluation of socialist national currencies, also adjusts, often

TABLE 5—ALTERNATIVE REGRESSIONS WITH OUTPUT PER WORKER, ADJUSTED FOR NFH AND LABOR QUALITY, AS DEPENDENT VARIABLE, 1975[a]

Regression	Scope of Dummy Variable	R^2	Regression Coefficient A	B	M	N	Q
II–1	Socialist Countries	.962	.567 (7.87)	.098 (3.98)	−.304 (5.45)		.047 (1.14)
II–2	Socialist Countries	.971	.567 (7.87)	.098 (3.98)	−.350 (5.73)		.047 (1.14)
II–3	Socialist Countries	.971			−.283 (4.20)		
II–4a	USSR	.950	.637 (6.13)	.101 (3.56)	−.320 (3.89)		.063 (1.26)
II–4b	Hungary				−.324 (3.73)		
II–4c	Poland				−.228 (2.13)		
II–4d	Yugoslavia				−.208 (1.76)		
II–5	Socialist Countries	.958	.585 (7.04)	.098 (3.78)	−.304 (5.17)	−.014 (0.55)	.034 (0.69)
II–6	Socialist Countries	.954	.631 (6.76)	.115 (3.44)	.386 (5.00)		−.032 (0.54)

[a]See Table 4. In the case of the USSR, land as well as labor has been adjusted for quality.

speculatively, for possible divergencies between capital stocks and services. That computation too hardly improves the socialist performance.[15]

In I–4a to I–4d (Table 4), I again repeat the calculations made for I–1, but separate dummy variables are assigned to the different socialist countries. With this, R^2 and the results for capital and land are more or less comparable to what they were in I–1. In all cases, too, socialism still has a negative coefficient. The magnitude of the latter varies, the negatives obtained for Poland and Yugoslavia being smaller than those for Hungary and the USSR. With the marked reduction in degrees of freedom, however, significance levels for the socialist coefficients decline, and for Yugoslavia the negative coefficient is reliable at little better than the 10 percent level. Intriguing as such contrasts are, note that they are especially affected by data limitations.

None of the regressions considered thus far allows for the possibility that output per worker might be affected by differing opportunities to exploit economies of scale. In a further computation (Table 4, I–5), I try to test that possibility by introducing into (1) an additional term, $N \log P$, where P is population and N is a coefficient to be determined. Evidently N has the wrong sign and is not significant at any interesting level. Other results, though, are almost the same as in I–1.

I have referred to regressions where labor is adjusted for nonfarm hours. When allowance is made also for labor quality, and

[15]The calculation allows as before for an overestimation of socialist GDMP by 10 percent and an overestimation of socialist capital inputs by 20 percent. The latter distortion could again be due simply to overvaluation of socialist currencies, but possibly it reflects also some comparative superannuation of socialist capital assets, with a resultant impairment of services. Additionally, I allow, for all countries, as in fn. 12 above, for differences in hours, shift work, and unfinished construction. The dummy variable for socialism, again significant at the 1.0 percent level, is now of a magnitude, (−) .309, implying a 26.6 percent shortfall of socialist output per worker below WME levels. For the rest, the results are essentially as before, though the coefficient for capital per worker increases to .620.

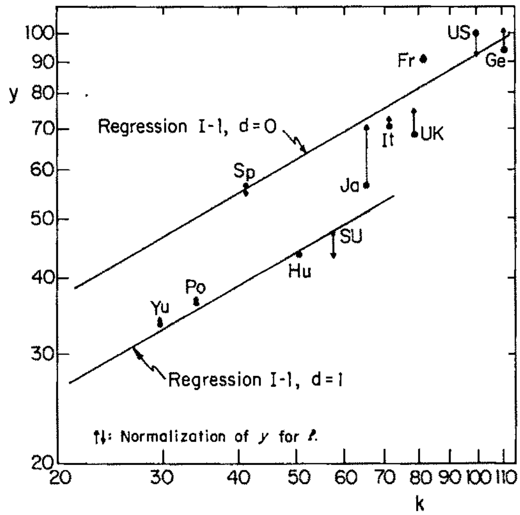


FIGURE 1. RELATION OF OUTPUT PER WORKER (y) TO CAPITAL PER WORKER (k) WITH LAND PER WORKER (l) ADJUSTED TO MEAN VALUE

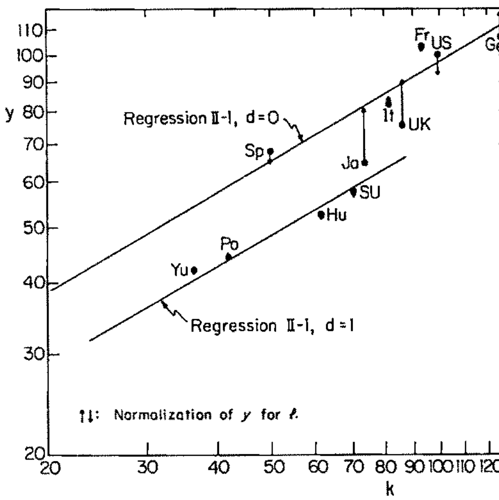


FIGURE 2. RELATION OF OUTPUT PER WORKER (y) AND CAPITAL PER WORKER (k), WITH LABOR ADJUSTED FOR QUALITY AND LAND PER WORKER (l) ADJUSTED TO MEAN VALUE

each computation is repeated, the results are much as before (Table 5). Socialism, however, now performs somewhat better than previously, for example, with labor adjusted for quality, as in II–1, output per worker under socialism is found to fall short of the WME level by 26.2 percent, instead of 29.6 percent as in the corresponding regression I–1, without such adjustment. The result is again significant at the 1.0 percent level. When separate dummy variables are assigned to the different socialist countries (II–4a to II–4d), they are again negative in all cases, but for Poland only at a 10 percent level and for Yugoslavia not even at that significance level.

With other country prices superseding those of the United States, GDMP per worker for WME countries other than the United States tend to compare less favorably with that country than they did before (Table 1). But, despite limitations of socialist prices, that familiar index number effect also holds for socialist countries. Hence, output per worker there continues to fall distinctly below WME levels. When (1) is applied, the dummy variable standing for socialism is again negative at the 1.0 percent level, and indicates a shortfall in socialist performance somewhat greater than that observed previously: by 33.6 percent when employment is adjusted for NFH and by 32.0 percent with further adjustment of employment for quality (Table 4, I–6; Table 5, II–6). Coefficients for capital and land per worker are much the same as before, though that for land now falls somewhat short of the 1.0 percent significance level.

By normalizing for land per worker, we can graph in two dimensions relations between output and capital per worker that have been described. With particular reference to regressions I–1 and II–1, the result (Figures 1 and 2) underlines what was already evident, that the calculated regression relations fit strikingly well the data they summarize. Interestingly, land, despite its relatively modest coefficient, turns out to be a consequential source of deviant productivity levels in some countries, especially the United Kingdom and Japan. The charts also illustrate, however, a feature alluded to at the outset: the lack of observations for WME countries at lower socialist levels of capital per worker. Granting the notably favorable T-statistics, my findings at such levels represent an extrapolation that remains to be tested by further observations for WME

countries. The same sort of caveat applies also where reference is to very high levels of capital per worker not yet experienced in the socialist world.[16]

IV. Conclusions

My principal conclusions are already fairly evident. Essentially, output per worker under socialism, as exemplified by the USSR, Hungary, Poland, and Yugoslavia, is found systematically to fall short of that in Western mixed-economy (WME) countries, such as the United States, the FRG, France, Italy, the United Kingdom, Japan, and Spain. I refer to output per worker in "material sectors," that is, the economy generally, exclusive of diverse services and housing; and to a residual discrepancy in that aspect that remains after allowance for differences in reproducible capital and land per worker.

As indicated by regression computations, the shortfall of socialist output per worker relative to that in WME countries is significant at the 1.0 percent level and of a magnitude ranging from 25 to 34 percent, the precise figure depending on whether allowance is made for differences in labor quality and possible data deficiencies, and also whether output and capital are valued in U.S. prices or, in each comparison with the United States, in the prices of the other country.

While comparative productivity is the immediate concern, the more ultimate one is comparative efficiency. The observed difference in performance between socialist and WME countries regarding output per worker should reflect any difference in efficiency between the two sorts of economic systems represented. In my regression calculations, however, I in effect normalize for conventional inputs, but fail to allow for a possible difference in technological knowledge. Reflecting variations in generation and borrowing of new knowledge, such a difference would not connote a difference in efficiency in the accepted sense understood in this essay. Once available in one country, however, technological knowledge seems soon to become available in another, and where it varies must often do so in dependence on a country's development stage. So far as it does, my calculations should in fact discount for it appropriately.

Thus, reproducible capital per worker, while representing the comparative inputs of the two factors, may also be viewed as an indicator of the development stage. Normalization for that coefficient, then, in effect normalizes as well for related differences in technological knowledge. The further result, also to the good, is normalization for the development stage more generally.[17]

The foregoing are results of regression computations where a single dummy variable stands for socialism in all four such countries considered. In further computations, where separate dummy variables are assigned to different socialist countries, the corresponding coefficients are still all negative. While with the reduction in degrees of freedom significance levels also deteriorate, the results appear to be consistent with those

[16]Judging from the graphs, output per worker under socialism not only falls short of but also may increase less rapidly with capital per worker than in WME countries.

[17]In seeking to acquire advanced technologies, the socialist countries in particular have doubtless been impeded by Western strategic controls, especially in some high-technology sectors. But the controls, directed primarily at military related technologies, have by all accounts been rather leaky. True, if the USSR is at all indicative, the socialist countries tend to lag behind the West in respect of civilian technologies, and perhaps more than might be expected of countries at their stage. But they have been able to obtain Western technologies on a vast scale, and if their technological levels tend to be unduly low, reasons enough can be found in the limitations of their own innovation processes. See my study (1983) and further works cited there on the Soviet experience in that sphere.

Economic advance results, among other things, from learning by doing. That is a process that takes time. May not our normalization, after all, be faulty so far as socialist countries, while accumulating capital relatively rapidly, have in effect had less time to learn? To a degree perhaps it is, but among WME countries, capital stock growth has been notably speedy in Japan and probably also in Spain. Yet output per worker in Japan falls but 2.1 percent below the norm for its input levels (under regression I–1). For Spain, too, there is a shortfall, but only of 3.3 percent.

of my earlier inquiry (1978, ch. 7) indicating that in respect of productivity the USSR was distinctly outclassed by a number of Western countries. My findings are still to be validated, however, by observations on WME performance at relatively early and on socialist performance at late development stages. The results for different socialist countries, in any event, differ to some extent, and divergencies in their economic working arrangements could be a cause. Without further inquiry, extension of our findings to other very different schemes could be hazardous.

I have assumed throughout that economic performance depends only on economic working arrangements. In complex ways still only imperfectly understood, such performance must also depend on the social system more generally—and that, Marx notwithstanding, can vary in a degree independently of the economic system. Quite similar economic systems might conceivably prevail and perform differently in different social contexts. The socialist schemes represented here thus might possibly perform better in another milieu. Such a qualified potentiality, however, is not what proponents of socialism have usually claimed for that system.

REFERENCES

Alton, Thad P. et al., *Economic Growth in Eastern Europe, 1965, 1970 and 1975–1982*, Occasional Paper 75, New York: L. W. International Financial Research, 1983.

Bergson, Abram, *Economics of Soviet Planning*, New Haven: Yale University Press, 1964.

______, *Planning and Productivity under Soviet Socialism*, New York: Columbia University Press, 1968.

______, "Index Numbers and the Computation of Factor Productivity," *Review of Income and Wealth*, September 1975, *21*, 259–78.

______, *Productivity and the Social System —The USSR and the West*, Cambridge: Harvard University Press, 1978.

______, "Technological Progress," in his and Herbert S. Levine, eds., *The Soviet Economy: Toward the Year 2000*, London: Allen and Unwin, 1983.

______, "On Soviet Real Investment Growth," *Soviet Studies*, forthcoming.

Cristensen, Laurits R., Cummings, Dianne and Jorgenson, Dale W., "Relative Productivity Levels, 1947–1973: An International Comparison," *European Economic Review*, May 1981, *16*, 61–94.

Denison, Edward F., *Accounting for Slower Economic Growth*, Washington: The Brookings Institution, 1979.

Edwards, Imogene, Hughes, Margaret and Noren, James, "U.S. and U.S.S.R. Comparisons of GNP," in Joint Economic Committee, U.S. Congress, *Soviet Economy in a Time of Change*, Vol. 1, Washington: USGPO, 1979.

Franz, A. et al., "Comparison of Prices and Levels of Gross Domestic Expenditures between Austria and Poland (APC), 1975 and 1978," *Statistical Journal of the United Nations Economic Commission for Europe*, 1982, 125–41.

Hibbert, J., Griffin, T. J. and Walker, R. L., "Development of Estimates of the Stock of Fixed Capital in the United Kingdom," *Review of Income and Wealth*, September 1977, *23*, 117–35.

Kravis, Irving B., Heston, Alan and Summers, Robert, *World Product and Income*, Baltimore: Johns Hopkins University Press: 1982.

Ofer, Gur and Vinokur, Aaron, *Private Sources of Income of the Soviet Urban Household*, R-2359 NA, Santa Monica: Rand Corporation, August 1980.

Powell, Raymond P., "The Soviet Capital Stock from Census to Census," *Soviet Studies*, January 1979, *31*, 56–75.

Schroeder, Gertrude E. and Edwards, Imogene, *Consumption in the USSR, An International Comparison*, Joint Economic Committee, U.S. Congress, Washington: USGPO, 1981.

European Economic Community Statistical Office, (Eurostat), *Comparison in Real Values of the Aggregates of ESA 1980*, Luxembourg: Eurostat, 1983.

Organization for Economic Cooperation and Development, (OECD), (1982a) *National Accounts*, Vol. 1, *1951–1980*, Paris: OECD, 1982.

——, (1982b) *Economic Surveys 1981–1982 Yugoslavia*, Paris: OECD, 1982.

Sovet Ekonomicheskoi Vzaimopomoshchi, (SEV), *Statisticheskii Ezhegodnik Stran-Chlenov Soveta Ekomicheskoi Vzaimopomoshchi 1981*, Moscow, 1981.

United Nations, *A System of National Accounts*, New York: United Nations, 1968.

——, Statistical Commission and Economic Commission for Europe, *International Comparison of Gross Domestic Product in Europe 1980*, New York: United Nations, 1985.

U.S. Department of Commerce, Bureau of Economic Analysis, *Fixed Reproducible Tangible Wealth in the United States, 1925–79*, Washington: USGPO, 1982.

VOL. 77 NO. 3 BERGSON: COMPARATIVE PRODUCTIVITY

The Adjustment of Expectations to a Change in Regime: A Study of the Founding of the Federal Reserve

By N. GREGORY MANKIW, JEFFREY A. MIRON, AND DAVID N. WEIL*

The founding of the Federal Reserve System in 1914 led to a substantial change in the behavior of nominal interest rates. We examine the timing of this change and the speed with which it was effected. We then use data on the term structure of interest rates to determine how expectations responded. Our results indicate that the change in policy regime was rapid and that individuals quickly understood the new environment they were facing.

How the economy reacts to a major change in the policy regime is an issue of widespread disagreement. At one extreme, some economists (for example, Thomas Sargent, 1982, 1983) suggest that if a change in regime is sufficiently credible, the economy will move quickly to the new rational expectations equilibrium. Yet others (John Taylor, 1975; Benjamin Friedman, 1979; Christopher Sims, 1982) argue that instant credibility is unlikely and that rational individuals should typically be expected to learn gradually about the new stochastic environment. This disagreement over how quickly economic agents perceive a change in their environment naturally leads to disagreement over the short-run impact of policy changes.

This paper is a case study of one particular change in regime—the introduction of the Federal Reserve System at the end of 1914. We use data on the term structure of interest rates to estimate how quickly individuals came to understand the new stochastic environment in which they were operating. Since long-term interest rates in part reflect expectations of future short-term interest rates, term structure data allow us to infer how expectations adapted to this change in regime.

In Section I we provide a brief historical overview of the introduction of the Federal Reserve System. Our emphasis in particular is on the prevailing view of the impact of the Fed prior to its beginning of operations. Such historical evidence is by its nature difficult to interpret and highly controvertible. Our reading of the historical record, however, is that observers during 1914 expected the Fed to effect a major change in the economic forces determining interest rates.

We document in Section II that a substantial change in the stochastic process of short-term interest rates did indeed occur. In the period from 1890 to 1910, short rates were quickly mean-reverting and highly seasonal. By contrast, in the period from 1920 to 1933, short rates were much more persistent; indeed, they were close to a random walk. There is little doubt that there was a major change in the stochastic process generating interest rates.

In Section III we examine the relation between long-term (six-month) and short-term (three-month) interest rates. Since the long rate incorporates an expectation of a future short rate, a change in the stochastic process generating short rates should alter the relation between long and short rates. In other words, as Robert Lucas's (1976) critique suggests, the parameters of traditional term structure equations relating long rates to short rates (for example, Franco Modigliani and Richard Sutch, 1966) should not remain invariant across regimes. In par-

*Departments of Economics, Harvard University; University of Michigan, Ann Arbor, MI 48109; and Harvard University, Cambridge, MA 02138, respectively. We are grateful to Barry Eichengreen, Milton Friedman, James Poterba, Angelo Melino, and an anonymous referee for comments. Mankiw received financial support from National Science Foundation grant no. SES-8520044.

ticular, since shocks to the short rate were less persistent in the 1890–1910 period than in the 1920–33 period, the long rate should be less responsive to the short rate in the earlier period. We find that the relation between six-month and three-month rates did in fact change in the way suggested by expectations-based theories of the term structure.

We examine in Section IV the timing of the change in regime. Using switching-regression techniques, we estimate that the most likely date for the change in the stochastic process of the short rate is between December 1914 and March 1915. This estimate, which uses only interest rate data, coincides almost exactly with the date at which the Federal Reserve began operation. We consider the possibility that the change in regime was gradual, but find instead that it occurred essentially all at once.

In Section V we study how quickly financial market participants perceived the change in regime. Our inferences are based on the premise that long-term interest rates depend on individuals' *perception* of the stochastic process the short rate is following. If there was a substantial lag in individuals' recognition of the change in their environment, then the relation between long rates and short rates should have changed long after the change in regime itself took place. By contrast, we find that the change in the relation between the six-month rate and the three-month rate roughly coincided with the change in regime. This finding suggests that financial market participants quickly understood the stochastic processes generated by the new policy regime and that, at least for this historical episode, the convergence to the new rational expectations equilibrium was quite rapid.

We conclude in Section VI. The evidence from the founding of the Fed suggests that a major change in a policy regime, backed with the establishment of new and powerful institutions, can be understood very quickly by financial market participants. It would of course be imprudent to extrapolate directly this single historical episode to the evaluation of other sorts of policy proposals. This episode does illustrate, however, the potential for rapid adjustment of agents' expectations in the face of substantial and widely believed changes in the continuing policy rule.

I. Historical Overview

The year 1914 witnessed two crucial events in the world of finance:[1] the creation of an important new institution, the Federal Reserve System, and the elimination of an old one, the classical gold standard.[2] In the sections that follow, we provide econometric evidence that there was a substantial change in regime and that this change was understood by financial market participants at the time. Our goal in this section is to show that such a conclusion is historically plausible; indeed, it is suggested by the literature of the time. After describing briefly the events surrounding the passage of the Federal Reserve Act and the opening of the Reserve Banks, we show that the relevant economic actors were aware that a regime change was taking place and had a rough idea of how the new regime would differ from the old.

The proximate cause of the founding of the Fed was the financial panic of 1907, which severely disrupted the economy and was widely blamed for the 1907–08 reces-

[1] The year 1914 also saw the outbreak of World War I. Our estimates of the stochastic process followed by the short-term interest rate indicate that the short rate followed essentially the same process in the 1915–18 period as in the 1919–33 period. It appears, therefore, that the war was not itself the major factor in the regime change examined here. Truman Clark (1986) has recently called into question whether the change in the behavior of interest rates at this time was due to the founding of the Federal Reserve, noting that a similar change took place in other countries as well. Clark provides no alternative explanation, however. While our econometric results below point to the founding of the Fed rather than the abandonment of the gold standard as the likely cause of the regime change, our analysis of the adjustment of expectations does not rely on the Fed being the source of the change.

[2] The classical gold standard effectively came to an end at the outbreak of World War I at the beginning of August 1914. During the period 1919–31, most countries expected to return to a fully operational gold standard and several resumed specie payments for limited periods. Overall, however, the period was not very similar to the classical gold standard era.

sion. In 1908, Congress passed the Aldrich-Vreeland Act, the most important result of which was creation of the National Monetary Commission. This group of legislators, academics, and bankers published a report in 1910 that discussed in enormous detail the positive and negative features of the United States' and foreign financial systems; the report served as a major impetus to the founding of the Fed. The Federal Reserve Act passed into law on December 23, 1913. The presidents of the banks met for the first time in July of 1914, and discussed the organization the system would take; the banks officially opened for business on November 16, 1914.

It is hard to believe that any change of regime was more widely perceived than the founding of the Federal Reserve. Paul Warburg, a well-known investment banker and advocate of the creation of the Fed, specifically applied the metaphor of a change in political regime, calling the Fed's founding "the Fourth of July in the economic life of our nation."[3] *The New York Times* for November 16, 1914, editorialized that "the starting of the Federal Reserve system, although incompletely, opens a new era in which 'old statistics do not count'" (p. 8). We could not hope for a more precise description of how an economic actor should respond to structural change.

The precise manner in which the Fed would operate was of course not known by financial market participants. The discussion in the report of the National Monetary Commission, however, makes clear that at least one essential function of the Fed was to operate a discount mechanism that would provide credit in times of excess demand, thereby dampening interest rate fluctuations and decreasing the frequency of bank failures. The day before the opening of the Fed, Secretary of the Treasury William McAdoo announced:

> The opening of these banks marks a new era in the history of business and finance in this country. It is believed that they will put an end to the annual anxiety from which the country has suffered for the last generation about insufficient money and credit to move the crops each year, and will give such stability to the banking business that extreme fluctuations in interest rates and available credits which have characterized banking in the past will be destroyed permanently.[4]

The financial press also believed that the introduction of the Fed would initiate an "elastic" currency and credit system.[5] No longer would interest rates have to move over such a great range to match the supply and demand for credit.

The evidence indicates strongly that financial market participants understood the intentions of the new institution. What we are unable to extract from the historical record is whether businessmen at the time of the Fed's founding expected it to accomplish its assigned tasks, or, alternatively, how long they expected the Fed would take to reach full operation. We can determine, however, that within a year of the opening of the Fed, popular opinion was that, as far as stabilization of the credit market was concerned, the Fed had accomplished all that it had set out to do. "What has thus far been done has been effectual in rendering stable and more uniform rates of discount prevalent throughout the country," wrote "Washington Notes" in the *Journal of Political Economy* (1915, p. 994; no author listed). On the subject of whether the Fed was wholly responsible for the year of ease in the credit markets that had followed its founding, *The*

[3] *Literary Digest*, November 27, 1915, quoting Warburg at the time of the founding.

[4] *The New York Times*, November 16, 1914, p. 1.

[5] *The Wall Street Journal* wrote, "The periodical convulsions in the money market for some time past had indicated clearly that there was something wrong with the currency medium of exchange of the country which was shown to be the lack of elasticity of circulation" (November 16, 1914, p. 1). *The New York Times* wrote, "When the new regime is fully operative, the currency volume will rise and fall with bank deposits, which will rise and fall with the course of trade" (November 16, 1914, p. 8).

New York Times wrote:

> Few will contend that the favorable progress of the year is altogether due to the betterment of the conditions of banking and of commercial credit through the operation of the Reserve system. Fewer still will contend that the system did not reenforce the forces making for recovery in ways that hardly anybody foresaw. No doubt the extremely easy money market assisted, but the money market would hardly have been so easy without the certainty that there would be no currency-scarcity under the Federal system.
> [November 17, 1915, p. 10]

II. The Stochastic Process of the Short Rate

The historical evidence presented above suggests that the behavior of short-term interest rates was a key feature of the change in regime associated with the founding of the Federal Reserve System. It is therefore natural to focus on this variable when studying the transition from the old regime to the new one.[6] The interest rate series that we examine here is the three-month time loan rate available at New York City banks for the first week of each month during the period from 1890 to 1933.[7] New York was already the major financial center of the country at this time. As John James (1978, pp. 61–64) reports, most loans in bank portfolios were short term and most loans in New York were fixed maturity. We are thus examining here the rates on an important form of short-term commercial credit. Since there was no significant Treasury bill market until the early 1930's, it is one of the principal short-term rates in the economy.

Table 1 shows the autocorrelations of the short rate during two different sample periods.[8] The first ends clearly before the changes that led to the new regime, while the second begins several years after the changes had occurred (as well as after the end of World War I). We present the autocorrelations for both the level of the rate and its first difference. The standard deviation of the short rate, both in levels and first differences, is provided at the bottom of the table.

TABLE 1—AUTOCORRELATIONS OF THE SHORT RATE

	1891–1910		1921–33	
	Level	Change	Level	Change
First	0.75	−0.18	0.95	0.03
Second	0.60	0.12	0.89	0.03
Third	0.39	−0.21	0.84	−0.10
Fourth	0.28	−0.04	0.79	0.09
Fifth	0.19	−0.05	0.74	0.12
Sixth	0.12	−0.09	0.67	0.05
Seventh	0.10	−0.01	0.60	−0.03
Eighth	0.09	−0.09	0.54	0.02
Ninth	0.11	0.01	0.48	−0.09
Tenth	0.13	0.00	0.44	0.00
Eleventh	0.14	0.08	0.38	−0.01
Twelfth	0.13	0.13	0.33	0.05
Standard Deviation	1.54	1.08	1.94	0.51

Note: The approximate standard errors for the autocorrelations are 0.06 for the 1890–1910 sample and 0.08 for the 1921–33 sample.

For the 1891–1910 period, the first autocorrelation of the level of the short rate is 0.75, and the autocorrelations die out fairly quickly. Seven out of the first eight autocorrelations of the change in the short rate are negative, indicating that the short rate was at least partly mean-reverting. For the 1921–33 period, the first autocorrelation of the level is close to one and the autocorrelations die out very slowly. All the autocorrelations of the change in the short rate are small for this later period.

The regression results in Table 2 confirm the impressions given by Table 1. We show, for the two sample periods, regressions of

[6] Our focus here on the nominal short rate and the term structure of nominal interest rates is not meant to imply that real interest rates are unimportant. The expectations theory implies a change in the relation between long and short nominal rates even if, as Robert Shiller (1980) suggests, the stochastic process for real rates did not change.

[7] This data set is described in the Data Appendix and is examined in Mankiw and Miron (1986a).

[8] We end the second sample in 1933 because in that year the Glass-Steagall Act introduced a variety of banking regulations. The results would be essentially the same if we ended the second period before the beginning of the Great Depression in 1929.

TABLE 2—REGRESSION OF SHORT RATE ON LAGGED SHORT RATE[a]

	Dependent Variable: r_{t+1}			
	1890–1910	($T=252$)	1920–1933	($T=168$)
Constant	1.01	0.03	0.09	0.05
	(0.18)	(0.29)	(0.09)	(0.16)
r_t	0.75	0.77	0.97	0.98
	(0.04)	(0.04)	(0.02)	(0.02)
$D2$		1.05		0.34
		(0.29)		(0.19)
$D3$		0.90		−0.10
		(0.29)		(0.19)
$D4$		0.55		−0.16
		(0.29)		(0.19)
$D5$		0.60		−0.07
		(0.29)		(0.19)
$D6$		0.77		−0.06
		(0.29)		(0.19)
$D7$		1.10		0.17
		(0.29)		(0.19)
$D8$		1.44		0.16
		(0.29)		(0.19)
$D9$		1.30		0.12
		(0.29)		(0.19)
$D10$		1.46		−0.12
		(0.29)		(0.19)
$D11$		0.71		−0.12
		(0.29)		(0.19)
$D12$		1.10		0.17
		(0.29)		(0.19)
$\bar{R}^2$	0.57	0.62	0.94	0.94
$s.e.e.$	1.00	0.94	0.52	0.51
D-W	2.09	2.20	1.92	1.91

[a]Standard errors are shown in parentheses.

the short rate on its own lagged value, including and excluding seasonal dummies. In the earlier period, the coefficient on the lagged short rate is significantly less than one, again indicating that the short rate was mean-reverting. Also, the seasonal dummies enter strongly significantly in the first period.[9] In the later period, the coefficient on the lagged short rate is close to one and the seasonal dummy variables do not enter significantly, suggesting that the short rate is close to a random walk. These results demonstrate that the process for the short rate was very different after the founding of the Federal Reserve and the abandonment of the gold standard.

III. The Short-Rate Process and the Term Structure of Interest Rates

In this section we examine the implications of expectations-based theories of the term structure for a traditional term structure equation, such as that suggested by Modigliani and Sutch. As the Lucas critique suggests, one should not expect such an equation to remain invariant when there is a fundamental change in the stochastic process generating short rates. We show that the parameters of a reduced-form equation estimated over the two regimes considered in the previous section did in fact change in the way one would have predicted.

A. *Theory*

Let r_t be the three-month yield and R_t be the six-month yield. Consider a reduced-form equation relating the longer-term rate to the short rate:

$$R_t = \alpha + \beta r_t + \varepsilon_t, \tag{1}$$

where α and β are parameters and ε_t is a random error. Equation (1) is the simplest version of the Modigliani-Sutch equation. This sort of equation, often with additional lags, is used for policy analysis both in large-scale models such as the MPS model (as noted by Olivier Blanchard, 1984) and in smaller-scale simulation models (for example, Richard Clarida and B. Friedman, 1984).

Expectations-based theories of the term structure relate the long-term rate to current and expected future short-term rates. With monthly data,

$$R_t = \tfrac{1}{2}(r_t + E_t r_{t+3}) + \theta_t, \tag{2}$$

where E_t denotes the expectation conditional on information available at time t and θ_t denotes the term premium. On the basis of the evidence discussed above, let us suppose the short rate follows a first-order autore-

[9]The seasonal fluctuations in interest rates, which are not of primary importance for the issues we address in this paper, are discussed in Milton Friedman and Anna Schwartz (1963, pp. 292–96), Shiller (1980), Miron (1986), Clark (1986), and Mankiw and Miron (1986b).

gressive process.[10] That is, ignoring the constant and seasonal dummies for simplicity,

$$r_{t+1} = \rho r_t + \nu_{t+1}. \tag{3}$$

Equations (2) and (3) imply that

$$R_t = \tfrac{1}{2}(1+\rho^3) r_t + \theta_t. \tag{4}$$

The standard expectations theory of the term structure, which is the hypothesis that the term premium is constant, thus implies a restriction across equations (1) and (3). In particular, it implies that

$$\beta = \tfrac{1}{2}(1+\rho^3). \tag{5}$$

The more persistent are shocks to the short rate (higher ρ), the greater is the response of the long rate to the short rate (higher β).

If the term premium θ_t is constant through time, as the expectations theory assumes, then equation (4) has no error. More generally, however, if the term premium varies but is uncorrelated with the short rate, then equation (4) has an error but this error does not change the restriction in equation (5). Since the restriction in equation (5) is much more general than the expectations theory, the abundant evidence against the expectations theory (for example, Robert Shiller, John Campbell, and Kermit Schoenholtz, 1983; Mankiw and Miron, 1986a,b) is not directly relevant to this restriction.

Once one interprets the error in the Modigliani-Sutch equation as the term premium, however, there is no reason to suppose it is serially uncorrelated. Below we quasi-difference equation (1) to correct for serial correlation. As long as the term premium is uncorrelated with the short rate at leads and lags, the restriction in equation (5) continues to hold.

We can now see the implications of a change in the stochastic process generating the short rate. Since the dynamic process of the short rate (equation (3)) changed from 1890–1910 to 1920–33, there should have been a change in the parameter of the Modigliani-Sutch relation (equation (1)). In particular, since shocks to the short rate became more persistent, the long-term interest rate should have become more responsive to the short-term interest rate.

B. *Evidence*

Tables 3 and 4 present estimates of equation (1) for the two sample periods considered in Section II. In Table 3 we use the level of long and short rates, while in Table 4 we use quasi-differenced data in order to account for serial correlation. The filter we use is $(1-0.5\,L)$, which is suggested by the Durbin-Watson (*D-W*) statistic of the regression in levels and appears to leave the residual approximately serially uncorrelated. The coefficient estimates we obtain with quasi-differenced data are not qualitatively very different from those we obtain with the raw data. We hereafter restrict our attention to the results with quasi-differenced data.

These results show clearly the effects of regime changes predicted by Lucas. In particular, the relation between long rates and short rates changed when the process for short rates changed in the way that the expectations theory predicts. The coefficient in the Modigliani-Sutch regression increased from 0.47 to 0.93 between the two periods. At least by the time period covered in our second sample, agents had come to understand that a new, more persistent, process for the short rate was in effect, and they had altered their behavior accordingly.[11]

[10] The assumption implicit here is that individuals have no information in forecasting the short rate other than the variables included in this equation. This assumption is obviously a strong one and can only be justified as an approximation. One test is to include the long rate in the forecasting equation, since the long rate would reflect any additional information on the future short rate. For the 1890–1910 period, the long-rate coefficient is statistically significant but the improvement in fit is very small: the standard error of estimate falls by only .027 (2.7 basis points). For the 1920–33 period, the long-rate coefficient is not statistically significant. Hence, the assumption that agents have little information additional to that in our posited forecasting equation appears empirically plausible.

[11] Stanley Fischer writes,

"It is indeed remarkable that the Lucas policy evaluation critique has triumphed without any detailed

TABLE 3—REGRESSION OF LONG RATE ON SHORT RATE[a]

	Dependent Variable: R_t			
	1890–1910	$(T=252)$	1920–1933	$(T=168)$
Constant	2.05	1.91	0.37	0.32
	(0.09)	(0.15)	(0.03)	(0.05)
r_t	0.59	0.60	0.94	0.94
	(0.02)	(0.02)	(0.01)	(0.01)
$D2$		−0.02		0.04
		(0.16)		(0.06)
$D3$		0.05		0.02
		(0.16)		(0.06)
$D4$		0.07		0.06
		(0.16)		(0.06)
$D5$		0.06		0.05
		(0.16)		(0.06)
$D6$		0.02		0.06
		(0.16)		(0.06)
$D7$		0.27		0.03
		(0.16)		(0.06)
$D8$		0.35		0.16
		(0.16)		(0.06)
$D9$		0.20		0.11
		(0.16)		(0.06)
$D10$		0.23		0.05
		(0.16)		(0.06)
$D11$		−0.08		0.01
		(0.16)		(0.06)
$D12$		−0.01		−0.03
		(0.16)		(0.06)
$\bar{R}^2$	0.76	0.76	0.99	0.99
s.e.e	0.51	0.50	0.17	0.17
D-W	1.11	1.14	1.08	1.07

[a]See Table 2.

TABLE 4—REGRESSION OF LONG RATE ON SHORT RATE: QUASI DIFFERENCED[a]

	Dependent Variable: $(1-0.5\ L)R_t$			
	1890–1910	$(T=251)$	1920–33	$(T=168)$
Constant	1.25	1.24	0.19	0.18
	(0.06)	(0.12)	(0.03)	(0.05)
$(1-0.5\ L)r_t$	0.48	0.47	0.94	0.93
	(0.03)	(0.03)	(0.01)	(0.01)
$D2$		−0.13		0.03
		(0.14)		(0.06)
$D3$		0.05		0.00
		(0.14)		(0.06)
$D4$		0.01		0.04
		(0.14)		(0.06)
$D5$		−0.04		0.01
		(0.14)		(0.06)
$D6$		−0.09		0.03
		(0.14)		(0.06)
$D7$		0.20		−0.01
		(0.14)		(0.06)
$D8$		0.20		0.14
		(0.14)		(0.06)
$D9$		0.06		0.03
		(0.14)		(0.06)
$D10$		0.17		−0.01
		(0.14)		(0.06)
$D11$		−0.12		−0.02
		(0.14)		(0.06)
$D12$		0.02		−0.04
		(0.14)		(0.06)
$\bar{R}^2$	0.58	0.58	0.98	0.98
s.e.e.	0.44	0.44	0.15	0.15
D-W	2.10	2.09	2.20	2.22

[a]See Table 2.

The results, however, are not completely consistent with the simple theory discussed above. While the sort of parameter drift observed is in line with that predicted by theory, the point estimates of the coefficient in the Modigliani-Sutch equation are somewhat different than predicted. The short-rate equation in Table 2 predicts a coefficient of 0.73 for the 1890–1910 period and 0.97 for the 1920–33 period, in contrast to the actual estimates of 0.47 and 0.93. Thus, for the earlier period, the point estimate is quite different from what the theory predicts.

Table 5 presents joint estimates of the two equations imposing the cross-equation restriction in equation (5). The estimate of the parameter in the Modigliani-Sutch equation is 0.61 for the 1890–1910 period and 0.94 for the 1920–33 period. Not surprisingly, these estimates are between those in Table 4 and those implied by Table 2. A formal likelihood ratio test of the cross-equation restriction between the short-rate equation and the Modigliani-Sutch equation rejects that restriction for the 1890–1910 period, but not for the 1920–33 period.[12]

empirical support beyond Lucas's assertion that macroeconometric models in the 1960s all predicted too little inflation in the 1970s. The general point made by the critique is correct and was known before it was so eloquently and forcefully propounded by Lucas. That the point has been empirically relevant, however, is something that should have been demonstrated rather than asserted" (1983, p. 271).

The evidence from the founding of the Fed provides such a demonstration.

[12]Under the assumption that the error in the Modigliani-Sutch equation is the term premium and independent of the error in the short-rate equation, the joint

TABLE 5—JOINT ESTIMATION IMPOSING CROSS-EQUATION RESTRICTION[a]

	1890–1910	($T=251$)	1920–33	($T=168$)
Dependent Variable	r_{t+1}	$(1-0.5\ L)R_t$	r_{t+1}	$(1-0.5\ L)R_t$
Constant	0.84	0.90	0.03	0.17
	(0.40)	(0.17)	(0.14)	(0.05)
r_t	0.61		0.96	
	(0.03)		(0.01)	
$(1-0.5\ L)r_t$		0.61		0.94
		(0.02)		(0.01)
$D2$	0.83	0.02	0.45	0.04
	(0.41)	(0.20)	(0.18)	(0.09)
$D3$	0.72	0.09	0.02	0.00
	(0.52)	(0.20)	(0.20)	(0.08)
$D4$	0.37	0.07	−0.05	0.05
	(0.44)	(0.22)	(0.24)	(0.06)
$D5$	0.37	0.07	0.03	0.01
	(0.46)	(0.23)	(0.22)	(0.06)
$D6$	0.50	0.02	0.04	0.03
	(0.43)	(0.20)	(0.22)	(0.07)
$D7$	0.84	0.30	0.27	−0.01
	(0.46)	(0.21)	(0.21)	(0.08)
$D8$	1.22	0.25	0.26	0.14
	(0.41)	(0.23)	(0.22)	(0.06)
$D9$	1.19	0.05	0.23	0.03
	(0.42)	(0.22)	(0.20)	(0.07)
$D10$	1.39	0.15	−0.02	−0.01
	(0.40)	(0.22)	(0.18)	(0.06)
$D11$	0.70	−0.17	−0.02	−0.02
	(0.39)	(0.19)	(0.22)	(0.06)
$D12$	1.03	0.06	0.28	−0.04
	(0.42)	(0.20)	(0.30)	(0.06)
$\bar{R}^2$	0.60	0.54	0.94	0.98
s.e.e	0.99	0.49	0.51	0.16
D-W	1.75	2.39	1.92	2.23
Likelihood Ratio Test $-\chi^2(1)$	39.8		0.8	

[a]See Table 2.

This statistical rejection of the cross-equation restriction appears attributable to the assumption that the term premium is uncorrelated with the short-term interest rate. To illustrate directly the covariation between the term premium and the short rate, we can regress the excess holding return on long bonds, $(R_t-0.5(r_t+r_{t+3}))$, on the short rate, r_t, adjusting the standard errors for the moving average residual. The coefficient on the short rate is $-.11$ with a t-statistic of 1.84 in the 1890–1910 period and -0.1 with a t-statistic of 0.35 in the later period. Hence, covariation between the term premium and the short rate appears to account for the statistical rejection in the early period.[13] While this covariation invalidates the cross-equation restriction in equation (5), a more persistent short rate (higher ρ) nonetheless

log likelihood is the sum of the two individual log likelihoods. We maximize the joint log likelihood by numerical optimization. We do not impose here cross-equation restrictions on the month dummies, which allows for the possibility of a seasonal term premium.

[13]Measurement error in the short rate is observationally equivalent to a negative covariation between the term premium and the short rate. While there is clearly some measurement error in these data, since the interest rates are the midpoint of a reported range of typically 12.5–25 basis points, we suspect that the measurement error is not sufficiently great to explain the results reported in the text.

leads, *ceteris paribus*, to a more responsive long rate (higher β). It is in this weaker sense that the evidence is consistent with the theory presented above.

IV. The Timing of the Change in Regime

In this section we try to pin down the timing of the change in the stochastic process for the three-month interest rate. We begin by determining the most likely date for the change in regime, conditional on the assumption that the change occurred all at once. We then consider the possibility that the change in regime occurred gradually over time.

A. *Step Switching*

Suppose that the process for the short rate obeyed

$$r_{t+1} = \kappa_o + \rho_o r_t + \nu_{t+1}, \qquad t = 1,\ldots,T_s - 1$$

$$r_{t+1} = \kappa_n + \rho_n r_t + \nu_{t+1}, \qquad t = T_s,\ldots,T$$

where T_s is the switch date (the first period of the new regime). Our goal is to estimate T_s. The procedure we use is the maximum likelihood procedure suggested by Stephen Goldfeld and Richard Quandt (1976) and recently applied by John Huizinga and Frederic Mishkin (1986) to the stochastic process followed by real interest rates. Assuming normal errors, the log likelihood function for this model is

$$\log L = -\frac{T}{2}\log(2\pi) - (T_s - 1)\log(\sigma_o^2) - (T - T_s + 1)\log(\sigma_n^2) - \frac{1}{2}\sum_{t=1}^{T_s-1}\left(\frac{\nu_{t+1}^2}{\sigma_o^2}\right) - \frac{1}{2}\sum_{t=T_s}^{T}\left(\frac{\nu_{t+1}^2}{\sigma_n^2}\right)$$

where σ_o^2 and σ_n^2 are the error variances in the old and new regimes. We can determine the maximum likelihood value for T_s by computing the maximum likelihood estimates of the parameters for all possible T_s's and then choosing the value of T_s with the maximum likelihood.

Table 6 shows the log likelihood of various possible switch dates around the maximum likelihood switch date.[14] According to these results, the most likely date of the new regime is December 1914 when month dummies are excluded, but February 1915 when month dummies are included. Remember that the Federal Reserve System opened for operation on November 16, 1914. This econometric estimate of the date of the new regime is thus very close to the date an historical account would suggest.

To judge the degree of confidence one should have in these point estimates of the date the new regime began, we calculate the posterior odds ratio for alternative switch dates. If one has diffuse priors (i.e., one considers all possible switch dates equally likely), then the ratio of the likelihood values for different switch dates produces the posterior odds ratio. The posterior odds ratio is the ratio of subjective probabilities of different switch dates conditioning on the data.[15]

Table 6 shows, for a range of possible switch dates, the posterior odds ratio of that date as a switch date compared to the maximum likelihood date. The months from December 1914 to March 1915 are all highly probable as the date of the regime change. The relative odds for the dates before December 1914 or after May 1915, however, are extremely low. Hence, although we cannot be certain of the exact date of the switch, we can conclude with a high degree of confidence that the date for the switch was

[14] We have searched over all possible switch dates 1890–1933, but only report values around the global maximum. Since the coefficient estimates are essentially the same as those in Table 2, we do not report them here.

[15] We view this posterior odds ratio as a simple metric for judging how flat or steep is the likelihood function. Note that for each switch date, the remaining parameters are chosen to maximize the likelihood. An alternative calculation (see, for example, Donald Holbert, 1982) would be to posit a prior joint distribution over all the parameters, to use the likelihood function to yield a posterior joint distribution over all the parameters, and then to integrate out the remaining parameters, to produce the posterior marginal distribution for the switch date. In our application, since the most likely values of the remaining parameters vary very little over plausible switch dates, we believe this latter calculation would produce similar conclusions.

TABLE 6—SWITCH DATE FOR SHORT-RATE EQUATION
($r_{t+1} = \kappa + \rho r_t$)

Date	Excluding Month Dummies		Including Month Dummies	
	$-\log L$	Posterior Odds Ratio	$-\log L$	Posterior Odds Ratio
1914:1	613.2	.000	576.3	.000
2	611.3	.000	573.8	.000
3	611.8	.000	574.1	.000
4	612.2	.000	574.6	.000
5	612.7	.000	574.9	.000
6	612.9	.000	575.2	.000
7	613.0	.000	575.2	.000
8	583.1	.005	546.2	.004
9	582.3	.011	545.0	.013
10	582.7	.007	545.1	.012
11	583.0	.006	545.6	.007
12	577.8	1.000	540.9	.803
1915:1	578.1	.741	540.9	.741
2	578.0	.819	540.6	1.000
3	578.8	.368	541.1	.631
4	579.5	.183	541.8	.304
5	580.2	.091	542.5	.160
6	581.0	.041	543.1	.084
7	581.7	.020	543.7	.045
8	582.3	.011	544.5	.021
9	583.0	.006	545.4	.009
10	583.7	.003	546.3	.004
11	584.4	.001	547.3	.001
12	585.1	.001	548.0	.001

Note: $\log L$ is the log of the likelihood function. The posterior odds ratio is the probability that the switch occurred at that date relative to the probability that the switch occurred at the date with the highest likelihood; this calculation is based on the estimated likelihood value and diffuse priors.

within a few months after the beginning of the Federal Reserve System.

Since the posterior odds ratio for any potential switch date before December 1914 is very low, the change in the stochastic process for short rates is more likely attributable to the founding of the Fed than to the abandonment of the gold standard.[16] The gold standard was suspended at the outbreak of World War I in August 1914. The results in Table 6 indicate that the months between the beginning of the war and the introduction of the Fed are more consistent with the old regime than with the new regime. A casual examination of the data easily explains this result. Between November 1914 and December 1914, the short-term interest rate fell from 6 to $4\frac{1}{8}$ percent. If the new (random walk) regime had already been in effect, such an event would have been very unusual: it would have required approximately a four standard deviation shock. Under the old (mean-reverting) regime, such an event was much less atypical: it required approximately a one-standard-deviation shock. Hence, these data imply that it is very unlikely that the new regime began before December 1914.[17]

[16]We do not intend to suggest that the abandonment of the gold standard was completely irrelevant. If the gold standard had continued in effect, the Fed may have been less able to affect nominal interest rates.

[17]If the single observation of the November–December drop in the short rate is excluded, we are unable to distinguish between the abandonment of the gold standard and the founding of the Fed as the cause of the regime change.

B. *Logistic Switching*

Our second procedure for determining the timing of the change in the process for short rates is to estimate a time-varying parameter model that allows the coefficients of the short-rate equation to change gradually over time, rather than moving instantaneously from the old to the new values as in the switching regression above. Specifically, we assume that the parameters of the short-rate equation follow a logistic curve. That is, the short-rate process is

$$r_{t+1} = \kappa_t + \rho_t r_t + \nu_{t+1},$$

while the parameters for this process change as

$$\kappa_t = (1 - L(t))\kappa_o + L(t)\kappa_n,$$

$$\rho_t = (1 - L(t))\rho_o + L(t)\rho_n,$$

$$\sigma_t^2 = (1 - L(t))^2\sigma_o^2 + L(t)^2\sigma_n^2,$$

where $L(t) = e^{\alpha+\delta t}/(1 + e^{\alpha+\delta t})$. All the parameters of the short-rate process adjust continuously together.

The parameters α and δ determine when the regime change occurs. In particular, at $t = -\alpha/\delta$, $L(t) = 1/2$ and the logistic curve has its inflection. At this date, the short-rate process is an equal mix of the old and the new regimes.

The parameter δ determines the rate at which the parameters change from their old values to their new values. Since $L(t)$ reaches one only asymptotically, the parameters approach their new values asymptotically. To judge the speed of the change in regime, define the dates $t(1/4)$ and $t(3/4)$ implicitly as

$$L(t(1/4)) = 1/4; \; L(t(3/4)) = 3/4.$$

Then $t(3/4) - t(1/4)$ is the period of time it takes for the parameters to make one-half of the adjustment (from one-fourth new regime to three-fourths new regime). Straightforward algebra shows that

$$t(3/4) - t(1/4) = \log(9)/\delta.$$

Hence, the parameter δ is inversely related to the rate of adjustment between regimes. The limit of the logistic curve ($\delta \to \infty$) is the step function, so this time-varying parameter model includes our earlier model as an extreme case.

Table 7 presents results for the logistic time-varying parameter specification of the short-rate process. The parameters are estimated with maximum likelihood assuming normal error (see Goldfeld and Quandt). We estimate the short-rate process both excluding and including month dummies. To reduce the computational problem, when month dummies are included, their coefficients are set equal to the values estimated for the old and new regimes as presented in Table 2.

Since the rate of adjustment is the key parameter here, we present the results for various rates of adjustment, choosing the remaining parameters to maximize the likelihood function. For each rate of adjustment, we present the maximum likelihood switch date ($L(T_s) = 1/2$), the maximum likelihood value achievable with that rate of adjustment, and the posterior odds ratio for that rate of adjustment relative to the maximum likelihood rate of adjustment.

The results in Table 7 indicate that either the step function ($\delta = \infty$) or a very steep logistic curve has the highest likelihood value. Since the implied switch dates for these curves are in the first few months of 1915, these steep logistic curves closely approximate the step function considered above. The likelihoods of less steep logistic curves, however, are much lower. We can conclude with a high degree of confidence that most of the change in regime occurred in less than one year.

V. Learning about the Change in Regime

In Section III we demonstrated that, at least after a period of several years, agents had correctly responded to the new stochastic process for the short rate. Here we estimate how quickly this response occurred. As in our treatment of the short-rate process, we examine both step switching and logistic switching.

The relationship between long rates and short rates depends on agents' perception of their environment. Suppose, for example,

TABLE 7—LOGISTIC SWITCHING FOR THE SHORT-RATE EQUATION
$(r_{t+1} = \kappa + \rho r_t)$

Months for 1/2 of Switch (δ)		Switch Date	$-\log L$	Posterior Odds Ratio
Excluding Month Dummies				
0.0	(∞)	1914:12	577.8	1.000
1.0	(2.197)	1915:1	578.2	.670
2.0	(1.099)	1915:2	577.9	.905
3.0	(0.732)	1915:1	578.1	.741
6.0	(0.366)	1915:4	579.2	.247
12.0	(0.183)	1915:8	582.0	.015
24.0	(0.092)	1916:7	584.7	.001
36.0	(0.061)	1916:0	585.4	.001
48.0	(0.046)	1916:2	586.6	.000
60.0	(0.037)	1917:3	587.8	.000
Including Month Dummies				
0.0	(∞)	1915:2	542.8	.670
1.0	(2.197)	1915:2	542.4	1.000
2.0	(1.099)	1915:2	542.4	.990
3.0	(0.732)	1915:2	542.5	.896
6.0	(0.366)	1915:4	543.9	.230
12.0	(0.183)	1915:7	547.6	.006
24.0	(0.092)	1916:10	549.6	.001
36.0	(0.061)	1916:0	551.1	.000
48.0	(0.046)	1916:1	553.1	.000
60.0	(0.037)	1917:2	554.7	.000

Note: $\log L$ is the log of the likelihood function for the set of parameters that maximizes the likelihood for the value of δ. The posterior odds ratio is the probability of that value of δ relative to the probability of the value of δ with the highest likelihood; this calculation is based on the estimated likelihood value and diffuse priors.

that even after the stochastic process for the short rate had changed to the more persistent process, agents had believed that the old mean-reverting process for the short rate was still in effect. (Such a situation might arise if agents had applied standard regression techniques to recent data to estimate the short-rate process.) In this case, fluctuations in the short rate would have been perceived as more transitory than they truly were. The long rate, which depends on the expected short rate, would have responded to the short rate as under the old regime. In other words, if perceptions adjusted gradually to the new regime, then the change in the empirical relationship between long and short rates should lag the change in the short rate process.

A. *Step Switching*

Table 8 presents a log likelihood of the Modigliani-Sutch equation for a range of possible switch dates around the maximum likelihood date.[18] The maximum likelihood switch date is December 1914 when month dummies are excluded and October 1914 when month dummies are included. The posterior odds ratio of all dates from October 1914 to January 1915 are fairly high. We can state with a high degree of confidence that the Modigliani-Sutch equation changed within a few months of the date the process for the short rate changed, even though we cannot be confident about the exact date. The data strongly support the conclusion that agents quickly understood that the introduction of the Fed had changed the stochastic environment in which they were operating.

[18]The coefficient estimates are essentially the same as those in Table 4.

TABLE 8—SWITCH DATE FOR THE MODIGLIANI-SUTCH EQUATION
$(1-0.5\,L)R_t = \alpha + \beta(1-0.5\,L)r_t$

	Excluding Month Dummies		Including Month Dummies	
Date	$-\log L$	Posterior Odds Ratio	$-\log L$	Posterior Odds Ratio
1914:1	159.0	.000	144.3	.000
2	159.3	.000	144.7	.000
3	160.3	.000	145.4	.000
4	161.2	.000	146.3	.000
5	162.4	.000	147.4	.000
6	162.5	.000	147.4	.000
7	162.6	.000	147.6	.000
8	161.0	.000	145.7	.000
9	110.0	.000	89.4	.000
10	83.8	.549	62.4	1.000
11	83.6	.670	62.9	.589
12	83.2	1.000	62.4	.951
1915:1	84.9	.183	64.1	.177
2	85.5	.100	64.7	.096
3	86.8	.027	65.8	.034
4	87.9	.009	66.8	.012
5	88.3	.006	67.3	.007
6	89.2	.002	68.0	.004
7	90.1	.001	69.0	.001
8	91.4	.000	70.9	.000
9	91.6	.000	71.8	.000
10	93.1	.000	73.6	.000
11	94.5	.000	75.4	.000
12	95.3	.000	76.1	.000

Note: See Table 6.

B. *Logistic Switching*

We present estimates of the logistic model for the Modigliani-Sutch equation in Table 9.[19] Both excluding and including month dummies, the maximum likelihood estimate for the time it took for the parameters to move halfway is one month, and the implied switch date is November 1914. The posterior odds ratios presented in the table show that adjustment periods of several months are reasonably likely, but that an adjustment period of six months or longer is highly improbable.

This result, that the participants in financial markets reacted quickly and properly to the change in the stochastic process of the short rate within a few months, is striking. It is clear that agents could not have estimated the new process for the short rate in just a few months. Our results suggest, nonetheless, that they had a good understanding of exactly what the new regime would be like. This finding is particularly dramatic because the new regime was not the sort of event for which there were many past observations from which to draw inferences.

Indeed, the data are consistent with an even stronger conclusion. We can see from the results that the Modigliani-Sutch equation may have changed before the process for the short rate changed. This finding suggests that agents anticipated the effects of the introduction of the Fed and modified their behavior accordingly, even before the Fed actually existed. If agents knew in October that the process for short rates would change in December, then the long rate implied by the expectations theory should have incorporated this fact. As we discuss in Section I, the Act establishing the Fed was passed in 1913, and the announcement of

[19] We again reduce the computational problem by using the estimates in Table 4 for the month dummies.

TABLE 9—LOGISTIC SWITCHING FOR THE MODIGLIANI-SUTCH EQUATION
$(1-0.5\,L)R_t = \alpha + \beta(1-0.5\,L)r_t$

Months for 1/2 of Switch (δ)		Switch Date	$-\log L$	Posterior Odds Ratio
Excluding Month Dummies				
0.0	(∞)	1914:12	83.2	.741
1.0	(2.197)	1914:11	82.9	1.000
2.0	(1.099)	1914:12	83.8	.407
3.0	(0.732)	1915:1	84.3	.247
6.0	(0.366)	1915:3	85.8	.055
12.0	(0.183)	1915:5	89.9	.001
24.0	(0.092)	1915:8	96.7	.000
36.0	(0.061)	1916:11	95.2	.000
48.0	(0.046)	1916:10	96.5	.000
60.0	(0.037)	1916:12	98.5	.000
Including Month Dummies				
0.0	(∞)	1914:11	65.9	.538
1.0	(2.197)	1914:11	65.3	1.000
2.0	(1.099)	1914:12	66.3	.353
3.0	(0.732)	1915:1	66.9	.200
6.0	(0.366)	1915:3	68.7	.034
12.0	(0.183)	1915:5	73.8	.000
24.0	(0.092)	1915:7	83.2	.000
36.0	(0.061)	1917:1	84.1	.000
48.0	(0.046)	1917:1	85.2	.000
60.0	(0.037)	1916:11	87.3	.000

Note: See Table 7.

the opening of the Fed occurred in July 1914. Thus, as a matter of history, agents did know when the Fed would begin operations. It is not implausible that agents also understood in advance the impact the Fed would have on the pattern of interest rates.

VI. Conclusion

The picture that emerges from this study is that of a remarkably fast adjustment of expectations and behavior in the face of a major change in the economic policy regime. We of course cannot determine exactly the timing and rate of adjustment to the new regime. Nonetheless, it would be difficult to reconcile these data with the hypothesis that agents observed the new regime for many months before responding to it.

Several caveats are in order. First, by looking only at term structure data, we are able to examine only the expectations of a relatively small group: New York financiers and businessmen who participated in the time loan market. Indeed, it may not even be necessary that all members of this group held the correct expectation right away; arbitrage by a well-informed subset might have produced the results we find. One should be cautious in applying our findings to situations in which the relevant expectations are those of a larger or less sophisticated group of economic actors.

Second, the implications of the regime change that we study, at least for short-term credit markets, were not difficult to predict. Since interest rate stability was one of the announced targets of Fed policy, no one should have been surprised that the stochastic process of short rates did in fact change. In many other cases of regime changes, the crucial expectations are those of nontarget variables. In these cases, the relevant economic actors must have an implicit or explicit model of the economy, which complicates their problem of understanding the new regime.

Finally, we note that observers in 1914 could have had a high degree of confidence that the Federal Reserve System would function as had been announced in advance. There was only modest political opposition

TABLE A1—THREE-MONTH INTEREST RATE

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1890	6.000	3.500	5.000	4.500	4.500	5.000	5.000	5.000	6.000	6.000	6.000	6.000
1891	6.000	4.500	5.000	4.500	4.000	5.750	4.500	4.750	6.000	6.000	6.000	4.000
1892	4.000	3.250	3.750	3.500	2.500	2.500	3.000	2.500	3.500	4.000	5.500	5.000
1893	6.000	3.500	6.000	5.500	6.000	4.750	6.000	6.000	6.000	6.000	4.250	2.750
1894	3.000	3.000	2.500	2.500	2.000	2.000	2.000	2.500	2.500	2.000	2.000	2.750
1895	2.500	3.250	3.250	3.750	2.500	2.000	2.000	2.500	2.500	2.750	2.500	3.000
1896	6.000	6.000	3.500	4.000	3.000	3.000	3.500	5.000	6.000	6.000	6.000	3.500
1897	3.000	2.500	2.500	2.500	2.500	2.500	2.000	2.000	3.000	3.500	3.000	2.500
1898	3.000	2.500	4.000	5.000	5.000	3.000	2.500	2.500	3.750	2.750	2.500	3.000
1899	3.000	3.000	3.750	4.000	3.500	3.000	3.000	4.750	4.000	6.000	5.750	6.000
1900	6.000	4.000	4.500	4.000	3.000	3.000	3.250	3.500	3.500	5.000	4.750	4.500
1901	4.500	3.250	3.000	3.500	4.250	3.250	4.000	4.375	5.000	4.750	4.500	4.000
1902	5.250	4.500	4.000	4.250	4.500	4.500	4.500	4.500	5.750	6.250	6.000	6.000
1903	5.250	4.750	5.250	5.375	4.500	4.750	4.000	4.500	5.000	5.750	5.750	5.750
1904	4.750	4.125	3.125	3.000	2.500	2.000	2.375	2.000	2.500	3.500	3.750	4.000
1905	3.125	2.875	3.125	3.375	3.250	2.875	3.000	3.250	3.625	4.875	4.875	5.375
1906	5.875	4.625	5.625	5.500	5.750	4.875	4.750	4.500	7.750	6.000	6.750	8.000
1907	6.750	5.500	5.250	5.000	3.750	4.500	4.625	5.500	5.750	6.250	14.000	10.00
1908	10.00	3.500	3.500	3.000	2.375	2.500	2.125	2.750	2.125	2.625	3.375	2.875
1909	2.625	2.500	2.875	2.625	2.625	2.500	2.375	3.000	3.375	3.875	4.625	4.750
1910	4.500	3.750	3.500	4.000	4.250	3.625	3.625	3.875	4.125	4.688	5.125	4.000
1911	3.750	3.125	3.000	2.875	2.750	2.875	2.750	3.125	3.250	3.500	3.625	3.750
1912	3.375	2.750	3.063	3.625	3.250	3.125	3.250	3.875	5.000	5.500	6.000	6.250
1913	5.000	4.000	4.750	4.250	4.000	4.375	3.625	4.875	4.625	4.625	5.000	5.375
1914	4.750	3.125	3.125	2.750	2.875	2.250	2.875	8.000	7.000	6.500	6.000	4.125
1915	3.625	2.875	2.875	2.750	2.750	2.625	2.750	3.000	2.750	2.750	2.750	2.500
1916	2.750	2.750	2.875	2.875	2.875	2.875	3.875	3.375	3.125	3.375	3.250	4.125
1917	3.750	2.875	4.125	3.875	4.375	4.125	4.250	4.375	5.250	5.750	5.500	5.375
1918	5.625	5.625	6.000	6.000	6.000	5.875	5.625	5.875	6.000	6.000	6.000	5.875
1919	5.375	5.125	5.500	5.625	5.875	5.625	6.000	6.000	5.875	5.875	6.500	6.500
1920	7.000	8.250	8.500	8.000	8.250	8.000	8.250	8.625	8.750	7.750	8.000	7.125
1921	7.375	6.750	6.750	6.750	6.625	6.875	6.500	5.750	5.875	5.375	5.500	5.125
1922	5.000	4.750	4.875	4.500	4.250	4.125	4.125	3.875	4.375	4.625	4.875	5.000
1923	4.750	4.750	5.000	5.375	5.125	4.875	5.125	5.125	5.500	5.500	5.125	5.000
1924	5.000	4.625	4.875	4.375	4.375	3.875	2.875	2.625	3.000	2.875	3.000	3.250
1925	3.875	3.625	3.875	4.125	3.875	3.750	3.875	4.250	4.375	4.625	4.875	4.938
1926	4.875	4.625	4.875	4.625	4.000	4.125	4.125	4.500	4.875	5.063	4.750	4.625
1927	4.625	4.438	4.438	4.375	4.375	4.438	4.500	4.313	3.938	4.313	4.250	4.063
1928	4.188	4.438	4.563	4.625	4.938	5.750	5.875	6.250	6.500	7.250	6.875	7.250
1929	7.625	7.625	7.750	8.750	8.625	8.375	7.375	8.875	8.875	9.125	6.000	4.875
1930	4.875	4.750	4.500	4.125	3.625	3.125	2.750	2.625	2.625	2.375	2.375	2.125
1931	2.375	1.875	2.125	2.125	1.875	1.375	1.625	1.375	1.625	2.500	3.750	3.250
1932	3.500	3.625	3.375	2.875	1.875	1.500	1.500	1.375	1.375	1.125	0.500	0.500
1933	0.500	0.500	3.000	1.500	1.125	0.875	0.875	1.375	0.625	0.688	0.688	0.875

to the new institution and no apparent benefits to the Fed in not fulfilling the expectations it had created. Our study does not speak directly to the problem of achieving credibility for an optimal but time-inconsistent policy.

The primary implication of all these caveats is that many particular circumstances facilitated the rapid adjustment of expectations to the regime change studied here. We therefore cannot be certain whether this phenomenon is to be found more generally. But the creation of the Federal Reserve does illustrate the surprising speed with which financial market participants can at times respond to a major change in the economic policy regime.

DATA APPENDIX

The data used in this paper are the time loan rates available at New York banks during the first week of the month from 1890 to 1933. In 1910, the National Monetary Commission compiled these data from 1890

TABLE A2—SIX-MONTH INTEREST RATE

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1890	6.000	4.500	6.000	5.500	5.000	5.500	6.000	5.250	6.000	6.000	6.000	6.000
1891	6.000	5.000	5.000	5.000	5.250	6.000	5.750	6.000	6.000	6.000	6.000	4.750
1892	4.750	4.000	5.000	4.000	3.750	3.250	3.750	4.000	4.500	5.000	6.000	6.000
1893	6.000	4.000	6.000	6.000	6.000	5.500	6.000	6.000	6.000	6.000	5.000	2.750
1894	3.750	3.750	3.500	3.000	3.000	3.000	3.000	3.250	3.750	3.000	3.000	3.250
1895	3.250	4.000	4.250	4.500	3.250	2.750	2.750	2.875	2.875	3.750	3.750	4.250
1896	6.000	6.000	4.000	4.750	3.500	3.500	4.000	5.750	6.000	6.000	6.000	4.000
1897	3.500	3.000	3.000	3.500	3.000	3.000	3.000	3.000	3.750	4.750	3.750	3.500
1898	3.750	3.000	4.000	5.000	5.000	3.000	3.500	3.250	4.000	3.250	3.000	3.500
1899	3.000	3.000	3.750	4.250	3.875	3.500	3.500	4.750	4.750	6.000	6.000	6.000
1900	6.000	4.250	4.750	4.000	4.000	4.000	4.250	4.250	4.000	5.000	4.750	4.750
1901	4.500	3.500	3.500	3.750	4.750	4.000	4.500	4.750	4.750	4.750	4.500	4.375
1902	5.250	4.750	4.375	4.000	4.250	4.500	4.500	4.875	5.000	6.000	5.500	5.875
1903	5.375	4.750	5.250	5.375	4.750	5.250	5.000	5.500	5.750	5.750	5.750	5.750
1904	4.750	4.375	4.000	3.875	3.375	3.125	3.375	3.500	3.500	3.750	3.750	3.875
1905	3.375	3.125	3.500	3.625	3.625	3.500	3.625	3.750	4.125	4.625	5.125	5.000
1906	5.625	4.625	4.375	5.250	5.750	4.875	5.625	5.625	6.500	6.000	6.000	6.250
1907	6.250	5.625	5.625	5.250	4.500	4.750	5.750	6.125	6.000	6.250	6.000	7.000
1908	6.000	4.625	4.375	4.000	3.625	3.500	3.500	3.875	3.625	3.500	3.750	3.500
1909	3.375	3.000	3.125	3.000	2.875	3.125	3.375	3.875	3.875	4.250	4.375	4.375
1910	4.500	4.125	3.875	4.125	4.250	4.125	4.500	5.000	4.875	4.625	4.875	4.000
1911	3.875	3.625	3.375	3.125	3.000	3.375	3.563	3.938	3.875	3.875	3.625	3.750
1912	3.625	3.125	3.375	3.875	3.500	3.500	4.125	4.875	5.125	5.375	5.750	5.750
1913	4.750	4.375	4.750	4.250	4.375	5.375	5.375	5.875	5.000	4.750	4.875	4.875
1914	4.750	3.500	3.375	3.000	3.375	3.000	3.875	6.000	7.500	6.500	5.750	4.125
1915	3.875	3.250	3.250	3.250	3.250	3.125	3.125	3.500	3.125	3.000	3.125	2.750
1916	3.000	3.000	3.125	3.000	3.125	3.125	4.000	3.875	3.625	3.500	3.375	4.125
1917	3.750	3.125	4.125	4.125	4.625	4.625	4.625	4.625	5.375	4.750	5.625	5.625
1918	5.750	5.875	6.000	6.000	6.000	6.000	5.875	5.875	6.000	6.000	6.000	6.000
1919	5.750	5.250	5.625	5.625	5.750	5.625	6.000	6.000	5.875	5.875	6.500	6.500
1920	7.000	8.250	8.500	8.500	8.250	8.000	8.250	8.375	8.750	7.750	7.875	6.250
1921	7.125	6.625	6.625	6.625	6.500	6.625	6.250	5.875	5.875	5.625	5.500	5.125
1922	5.000	4.750	4.875	4.625	4.375	4.250	4.250	4.250	4.500	4.875	4.875	5.000
1923	4.750	4.750	5.000	5.375	5.375	5.000	5.125	5.125	5.500	5.500	5.125	5.000
1924	5.000	4.625	4.875	4.625	4.625	4.125	3.250	3.500	3.375	3.125	3.375	3.625
1925	3.875	3.875	4.250	4.250	3.875	3.875	3.938	4.563	4.625	4.750	4.875	4.938
1926	4.875	4.625	4.875	4.625	4.125	4.125	4.125	4.625	4.875	5.063	4.875	4.625
1927	4.625	4.500	4.438	4.438	4.438	4.438	4.563	4.500	4.313	4.313	4.313	4.188
1928	4.188	4.500	4.563	4.813	4.938	5.750	5.875	6.250	6.500	7.000	6.750	7.125
1929	7.625	7.625	7.750	8.500	8.500	8.375	7.625	8.875	8.875	9.125	5.875	4.875
1930	4.875	4.875	4.500	4.125	3.875	3.625	3.000	3.125	3.125	2.750	2.750	2.625
1931	2.875	2.375	2.625	2.375	2.375	1.875	1.875	1.875	1.875	2.750	3.750	3.250
1932	3.500	3.625	3.375	2.875	1.875	1.500	1.500	1.375	1.375	1.125	1.000	1.000
1933	0.875	0.875	3.000	1.875	1.250	1.250	1.125	1.750	1.125	0.875	0.688	1.000

to 1909 by tabulating them from the *Financial Review*. We updated these series using the *Review* and the *Commercial and Financial Chronicle*, which took over from the *Review* in 1921. The rates are reported as a range, which is typically 12.5 to 25 basis points in size. We use the midpoint of the range. Tables A1 and A2 report all the data used.

REFERENCES

Blanchard, Olivier J., "The Lucas Critique and the Volcker Deflation," *American Economic Review Proceedings*, May 1984, *74*, 211–15.

Clarida, Richard H. and Friedman, Benjamin M., "Why Have Short-Term Interest Rates Been So High?," *Brookings Papers on Economic Activity*, 2:1983, 553–78.

Clark, Truman, "Interest Rate Seasonals and the Federal Reserve," *Journal of Political Economy*, February 1986, *94*, 76–125.

Fischer, Stanley, "Comment," in James Tobin, ed., *Macroeconomics, Prices and Quantities*, Washington: The Brookings Institution, 1983, 267–75.

Friedman, Benjamin M., "Optimal Expectations and the Extreme Information Assumptions of 'Rational Expectations' Macromodels," *Journal of Monetary Economics*, January 1979, *5*, 23–42.

Friedman, Milton and Schwartz, Anna J., *A Monetary History of the United States, 1867–1960*, Princeton: Princeton University Press, 1963.

Goldfeld, Stephen M. and Quandt, Richard E., "Techniques for Estimating Switching Regressions," in their *Studies in Nonlinear Estimation*, Cambridge: Ballinger, 1976.

Holbert, Donald, "A Bayesian Analysis of a Switching Linear Model," *Journal of Econometrics*, May 1982, *19*, 77–87.

Huizinga, John and Mishkin, Frederic S., "Monetary Policy Regime Shifts and the Unusual Behavior of Real Interest Rates," in Karl Brunner and Alan Meltzer, eds., *Carnegie-Rochester Conference on Public Policy: The National Bureau Method, International Capital Mobility, and Other Essays*, Spring 1986, Vol. 24, 231–74.

James, John A., *Money and Capital Markets in Postbellum America*, Princeton: Princeton University Press, 1978.

Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Alan Meltzer, eds., *The Philips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conference on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

Mankiw, N. Gregory and Miron, Jeffrey A., (1986a) "The Changing Behavior of the Term Structure of Interest Rates," *Quarterly Journal of Economics*, May 1986, *101*, 211–28.

______ **and** ______, (1986b) "Seasonality, Band Spectrum Regression, and the Expectations Theory of the Term Structure of Interest Rates," unpublished paper, 1986.

Miron, Jeffrey A., "Financial Panics, the Seasonality of the Nominal Interest Rate, and the Founding of the Fed," *American Economic Review*, March 1986, *76*, 125–40.

Modigliani, Franco and Sutch, Richard, "Innovations in Interest Rate Policy," *American Economic Review Proceedings*, May 1966, *56*, 178–97.

Sargent, Thomas, J., "The Ends of Four Big Inflations," in Robert E. Hall, ed., *Inflation: Causes and Effects*, Chicago: University of Chicago Press, 1982 (reprinted in *Rational Expectations and Inflation*, New York: Harper & Row, 1986).

______, "Stopping Moderate Inflations: The Methods of Poincaré and Thatcher," in R. Dornbusch and M. H. Simonsen, eds., *Inflation, Debt, and Indexation*, Cambridge: MIT Press, 1983 (reprinted in *Rational Expectations and Inflation*, New York: Harper & Row, 1986).

Shiller, Robert J., "Can the Fed Control Real Interest Rates?," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.

______, **Campbell, John Y. and Schoenholtz, Kermit L.,** "Forward Rates and Future Policy: Interpreting the Term Structure of Interest Rates," *Brookings Papers on Economic Activity*, 1:1983, 173–217.

Sims, Christopher A., "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, 1:1982, 107–52.

Taylor, John B., "Monetary Policy During a Transition to Rational Expectations," *Journal of Political Economy*, October 1975, *83*, 1009–21.

Awarding Monopoly Franchises

By MICHAEL H. RIORDAN AND DAVID E. M. SAPPINGTON *

We explain how to award a monopoly franchise so as to maximize expected consumers' welfare. Potential producers initially possess imperfect private information about production cost. The franchise is awarded to the producer with the lowest expected costs, but prices exceed realized marginal costs. These ex post distortions foster more competitive bidding ex ante. The distortions for any bid-cost pair are invariant to the number of bidders (n), though expected distortions and profits decline with n.

Awarding a monopoly franchise is quite common in practice. For example, local governments select a single firm to provide cable television service within designated geographic regions; they also grant monopolies for such municipal services as garbage collection. At least since Harold Demsetz (1968), the economics literature has recognized that, when the production technology is characterized by increasing returns to scale, there are potential gains from awarding a monopoly franchise by a competitive bidding process. Yet formal investigations of how to award monopoly franchises are surprisingly rare. Our intent is to help fill this void.

We examine the optimal procedure by which to award a monopoly franchise under conditions of cost uncertainty and incomplete information about the production technology. The regulator's objective is to maximize the expected value of consumers' surplus net of transfer payments to the producer (i.e., the single firm that is awarded the franchise).[1] There are an exogenous number of risk-neutral firms that are potential producers of the regulated commodity. These firms initially have independent private information concerning the prospective (increasing returns) production technology. The producer is the only party that learns its actual production costs. The regulator and all firms are presumed to know consumer demand for the product in question and to have common prior beliefs about technologies.

Martin Loeb and Wesley Magat (1979) proposed one possible strategy for the regulator to pursue in such a setting. They noted that a subsidy equal to realized consumers' surplus would induce the producer to price at marginal cost. Furthermore, taking this subsidy scheme as given, they argued that competitive bidding for the monopoly franchise would eliminate any rents to the producer. Thus, though the regulator may have no knowledge whatsoever of the production technology, he can effect the outcome that he would enforce if he shared the firm's private cost information.

We find that the Loeb-Magat (L-M) scheme is optimal for a Bayesian regulator only in the limiting cases where bidders are identical or where there are an infinite number of bidders. More generally, the optimal

*Stanford University, Stanford, CA 94305 and Bell Communications Research, Morristown, NJ 07960, respectively. Riordan acknowledges financial support from the National Science Foundation through grant no. IST85-07300. The views expressed in this paper are not necessarily those of Bell Communications Research. We thank David Baron, Tom Holmes, Richard Kihlstrom, Roger Klein, Tom Palfrey, Joseph Stiglitz, Oliver Williamson, and seminar participants at Bell Communications, Carnegie-Mellon, Duke, Northwestern, and Stanford universities, and the universities of Pennsylvania, Rochester, and Southern California for helpful comments. We are particularly grateful to an anonymous referee for valuable insights.

[1] Our qualitative conclusions would be unaltered if the regulator's objective were to maximize a convex combination of consumers' surplus and profits, provided the weight on the former exceeds the weight on the latter.

scheme will not induce efficient pricing *ex post*. *Ex post* distortions (prices in excess of marginal cost) will generally be optimal to promote more competitive bidding *ex ante*. The distortions, which vary with the winning bid, affect the expected profitability of the franchise differentially according to the bidder's private information, and thus induce bidders to self-select. The increment to expected revenues generated at the bidding stage as a result of self-selection more than offsets the surplus foregone at the production stage.

The regulator's optimal scheme can be given an interpretation of "menu contracting." The regulator designs a menu of franchise contracts. Each contract defines allowed prices and net transfer payments as functions of the monopolist's reported marginal production cost; the net transfer payment is equal to an *ex post* production subsidy less an "up-front" franchise fee.[2] The contracts on the menu are rank ordered, and a "bid" is a commitment to produce according to the terms of a particular contract. The monopoly franchise is awarded to the firm that selects (or bids) the most highly ranked contract. As will be shown, more highly ranked contracts generally involve prices closer to marginal cost and more generous production subsidies; they also require the firm to pay a larger franchise fee.

The discussion proceeds as follows. Section I summarizes the essential elements of the analysis and presents a formal statement of the regulator's problem. Section II characterizes the solution to this problem and states our major conclusions, while Section III offers intuitive explanations. Section IV summarizes briefly, and draws some conclusions. The Appendix provides a proof of our main theorem.

Before proceeding to our formal model, we offer some further perspectives on our research. We view our analysis as a component of the larger problem of how to organize a natural monopoly. There are three important aspects of the problem: (*i*) the selection of the producer(s); (*ii*) the determination of how much of the commodity is produced, and (*iii*) the distribution of the surplus from production.[3] Until very recently, most studies have ignored the first aspect, even though it is of great importance when potential producers have different technological capabilities, as is presumed in our analysis. (The recent exceptions are the independent works of Jean-Jacques Laffont and Jean Tirole, 1985, and R. Preston McAfee and John McMillan, 1987a,b.)[4] We consider all three aspects of the problem, though we do assume that the franchise is awarded to a single private producer. Consequently, our analysis does not consider many alternative organizational modes, such as public ownership or *ex post* competition among producers (James Anton and Dennis Yao, 1987).[5]

Our formulation also is motivated in part by the criticism that the standard principal-agent framework presumes a situation of bilateral monopoly and yet grants the principal all of the bargaining power. An example is David Baron and Roger Myerson's (1982) analysis of regulating a monopolist with unknown costs.[6] Our analysis builds on theirs, and is concerned with how a bilateral monopoly at the production stage arises from competition at a previous bidding stage, in

[2] Positive net transfer payments might be raised via a fixed access charge in a two-part tariff (Riordan, 1984). In this case, the access charge is presumed to be sufficiently small so as not to affect demand.

[3] There are other aspects of the monopoly problem that we do not address, such as providing incentives for investment or innovation.

[4] These papers assume that realized production costs are observable, while the effort exerted by the producer to reduce production costs is unobservable. Thus, these papers presume an incentive problem caused by unobservable actions follows the initial problem caused by private information. Our focus is on two successive problems complicated by private information.

[5] For an overview of the choice of governance structure under conditions of imperfect information, see B. Caillaud et. al (1985) or Sappington and Joseph Stiglitz (1987).

[6] Their analysis can be given the normative interpretation of selecting a particular "interim efficient" allocation rule (Daniel Spulber, 1987).

which the principal appears naturally as a monopsonist.[7]

Finally, it is important to note that while our model explicitly allows for cost uncertainty at the time the monopoly franchise is awarded, it does not address some of the other important practical objections to monopoly franchising raised by Oliver Williamson (1976). In particular, consumer preferences are known to the regulator and the quality of the product is not an issue in our model. Furthermore, bidding is not repeated here, and both the regulator and the producer are able to costlessly commit to the terms of the franchise contract: thus, "renegotiation" is not an issue.[8] Finally, we abstract from the costs of writing complicated contracts.

I. The Regulator's Problem

We consider the task of a risk-neutral regulator who arranges for the provision of a good produced under increasing returns to scale. The production technology is known to be characterized by the cost function, $C(Q,c)=K+cQ$. Here, $C(\cdot)$ represents total production costs, $c \geq 0$ is the marginal cost of production, $Q \geq 0$ is the level of output, and $K \geq 0$ represents sunk, fixed cost of production. The regulator's goal is to maximize expected consumers' surplus, $W(Q)$, net of transfer payments to the producer. We presume a well-behaved demand curve for the product in question, so that $W(\cdot)$ is an increasing, strictly concave, and continuously differentiable function of Q. The corresponding inverse demand curve is denoted by $P(\cdot)$.[9,10]

The essence of the regulator's problem lies in the information structure of the environment. There are two key elements. First, the realized marginal production cost (c) is never observed by any party other than the producer, and the producer learns c only after being awarded the franchise and incurring the fixed costs of production. Second, before bidding occurs, the ($n \geq 1$) producers each observe an informative private signal, t, about prospective operating costs, c. Thus, all uncertainty concerns the marginal cost, c. For simplicity, we assume the fixed cost of production is the same for all firms, and is common knowledge.

To describe the information structure formally, let $F(c|t)$ represent the cumulative conditional probability distribution of c given the signal $t \in [0,1]$. The term $F_1(c|t)$ represents the associated conditional density function, and has strictly positive support on the interval $[\underline{c}, \bar{c}]$. To associate the prospect of low production costs with high realizations of t, let $F_2(c|t) \geq 0$, where the subscript 2 (here and throughout) denotes the derivative with respect to the second argument. Thus, t realizations close to unity are "good news" in the sense of Paul Milgrom (1981), while those closer to zero are "bad news." It is apparent that this is a fairly general construct. One special case of our analysis would have t inversely related to the expected value of marginal production costs. We will often refer to t_i as the ith firm's *valuation* of the franchise.

The vector of private signals $(t_1, \ldots, t_n)$ of bidders 1 through n, respectively, are taken to be independent draws of a random variable that is uniformly distributed on the interval $[0,1]$.[11] Moreover, note that the conditional beliefs, $F(c|t_i)$, of the firm that observes signal t_i do not depend on the

[7]Thus, we consider formally what Williamson (1985) terms the "Fudamental Transformation," whereby an environment that is initially competitive is transformed into a bilateral monopoly because of sunk idiosyncratic (i.e., relation-specific) investments.

[8]See our paper (1986a) for an extension of the present analysis to the case in which the intertemporal commitment abilities for both the regulator and firms are restricted.

[9]By definition, $W(Q)=\int_0^Q P(s)\,ds - P(Q)Q$.

[10]Though our analysis proceeds in terms of a regulator overseeing a public utility, other interpretations are possible. For example, the problem considered might be that of the Department of Defense procuring spare parts, or a downstream firm contracting for the supply of an idiosyncratic input.

[11]The assumption of a uniform distribution for the underlying random variable is without essential loss of generality, in that a straightforward transformation can allow for any smooth density function. The independence assumption, however, does matter. The case of correlated private signals is discussed briefly in Section IV.

signals of other bidders. Thus, in the vernacular of the auction literature, this is a private values model for which "winner's curse" considerations are not relevant.[12] The assumption of independent private signals is intended to focus attention on technological differences that are idiosyncratic to individual producers as opposed to technological characteristics that are common to all producers.[13]

The bidding process is captured formally as follows. Potential producers simultaneously announce their private valuations of the franchise (i.e., firm i reports $\hat{t}_i$, $i=1,\ldots,n$).[14] As noted above, an announcement might readily be interpreted as a summary projection of expected costs. The firm that reports the highest t realization (i.e., predicts the lowest production costs) is granted the sole right to produce. Thus, the highest reported t realization constitutes the winning bid. The unit price the producer is allowed to charge and the net transfer payment he receives for production depends on both the winning bid t and the level of marginal cost $\hat{c}$, subsequently reported by the producer. The established price, $p(\hat{c},\hat{t})$, generates a level of demand $Q(\hat{c},\hat{t})$, according to the known demand curve, $Q(\hat{c},\hat{t})=P^{-1}(p(\hat{c},\hat{t}))$. The associated net transfer to the producer is represented as $T(\hat{c},\hat{t})$, and can be interpreted as having two components. One is a *production subsidy*, $S(\hat{c},\hat{t})$, paid to the producer after production occurs. The other (negative) component is a franchise fee, $\phi(\hat{t})$, paid by the winning bidder immediately upon being awarded the franchise. Thus, $T(\hat{c},\hat{t})\equiv S(\hat{c},\hat{t})-\phi(\hat{t})$. The producer is obligated to serve all demand at the established price. The "menu contracting" interpretation of the procedure discussed in the introduction arises because prices and transfer payments can vary with the winning bid. A particular bid is thus equivalent to selecting a price-transfer schedule from the menu specified by the regulator. The schedule dictates the terms of compensation for the firm at the production stage.

The timing in the model is as follows. First, the firms (costlessly) acquire private signals about their respective technologies. Second, the regulator announces the terms of the bidding procedure (i.e., the menu of contracts). Third, each firm simultaneously announces a bid for the franchise. Next, the franchise is awarded to the highest bidder. The winning bidder then pays the requisite franchise fee, and losing bidders receive their (common) reservation profit level, normalized to zero. Next, the winning bidder incurs fixed costs, K, learns marginal production cost, and makes a report, $\hat{c}$, about the realized cost parameter. Next, the regulated price is established and production occurs to fulfill demand. Finally, sales revenues flow to the producer, who also receives the production subsidy specified in the franchise contract.

Some features of our formulation warrant additional discussion. First, as noted earlier, the right to produce is awarded to only a single bidder. Implicitly, we are assuming K is so large that it would be prohibitively costly to have more than one firm sink this cost.[15] Second, $W(Q)$ is assumed to be sufficiently large relative to $\bar{c}$ in the neighbor-

[12] For a discussion of these terms and an overview of the literature on bidding mechanisms, see Milgrom and Robert Weber (1982) and the references they cite. Note that our analysis differs from most analyses in the literature on auction design in that we allow the "auctioneer" to choose the specifications of the "object" that is being auctioned.

[13] Suppose each firm's technology for producing the good in question is entirely idiosyncratic, and is due to historically random influences such as economies of scope with other activities, ownership of scarce resources, past investments, learning by doing, etc. Suppose also that the (symmetric) residual uncertainty at the time of bidding only concerns factor prices that are never observed by the regulator. Then the private information across firms may be characterized as independent realizations of a random variable.

[14] By the Revelation Principle (Partha Dasgupta et al., 1970, or Roger Myerson, 1979), this representation is without loss of generality.

[15] Thomas Palfrey has pointed out to us that if $[\bar{c}-\underline{c}]$ were sufficiently large, the regulator might gain from contracting with a second supplier if the first reports a cost realization close to $\bar{c}$. To rule out this behavior by the regulator, we presume that $[\bar{c}-\underline{c}]$ is small relative to K, so that the redundant sunk costs incurred by a second producer outweigh the maximum potential gain to the regulator from a second "draw" of marginal production cost. Note also that, in practice, the time delay associated with contracting with a second potential producer may be inordinately costly.

hood of $Q=0$ that the regulator always will ensure a strictly positive level of output; in particular, there is no "entry fee" for the auction (John Riley and William Samuelson, 1981). These are maintained assumptions which influence our findings.

There are three other assumptions implicit in the formulation of the regulator's problem below. These are: 1) the franchise is awarded to the most efficient bidder (i.e., the firm with the highest t realization); 2) prices and subsidies for the producer depend only on his bid, and not those of the losing bidders; and 3) losing bidders neither make nor receive payments. These additional assumptions are without further loss of generality, given risk neutrality, independent valuations, and a regularity condition (RC) that is developed in Section II.[16]

The regulator's problem [RP] is written formally as follows, where the Revelation Principle justifies restricting attention to mechanisms which induce truthful reports.

$$\underset{Q(\cdot),S(\cdot)}{\text{Maximize}} \int_0^1 \int_{\underline{c}}^{\bar{c}} [W(Q(b,t)) - T(b,t)] \times nt^{n-1} F_1(b|t)\, db\, dt$$

subject to

$$(1)\quad V(t) \equiv [t]^{n-1} \times \int_{\underline{c}}^{\bar{c}} \{[P(Q(c,t)) - c]Q(c,t) - K + T(c,t)\} F_1(c|t)\, dc \geq 0 \quad \forall t \in [0,1],$$

$$(2)\quad V(t) \geq [\hat{t}]^{n-1} \times \int_{\underline{c}}^{\bar{c}} \{[P(Q(r(c),\hat{t})) - c]Q(r(c),\hat{t}) - K + T(r(c),\hat{t})\} F_1(c|t)\, dc \quad \forall r: [\underline{c},\bar{c}] \to [\underline{c},\bar{c}], \quad \forall t, \hat{t} \in [0,1],$$

$$(3)\quad [P(Q(c,t)) - c]Q(c,t) + T(c,t) \geq [P(Q(\hat{c},t)) - c]Q(\hat{c},t) + T(\hat{c},t) \quad \forall c, \hat{c} \in [\underline{c},\bar{c}], \quad \forall t \in [0,1].$$

[16]A proof of this conclusion is available upon request from the authors. The proof proceeds along the lines of Myerson (1981).

In the statement of [RP] and throughout the ensuing discussion, $p(c,t) \equiv P(Q(c,t))$ is the unit price charged by the producer and $T(c,t)$ is the associated net transfer payment to the producer when the realized production cost is c and the winning bid is t. $r(c)$ represents any reporting strategy the producer might adopt, specifying his production cost report as a function of the realized cost.

The maximand reflects the regulator's goal of maximizing the expected level of consumers' surplus net of transfer payments to the producer; nt^{n-1} is the density of the maximal order statistic of $(t_1,\ldots, t_n)$. The individual rationality constraints (1) guarantee each bidder nonnegative expected profit. Notice that expected profit for a bidder with valuation t conditional on having won the franchise is discounted by the probability t^{n-1} that he wins the franchise (which is the probability that the other $n-1$ bidders all truthfully announce lower valuations). The truthful bidding contraints (2) guarantee that each firm anticipates greater profit if he announces his private information (t) truthfully (presuming others do the same) than if he misrepresents his valuation of the franchise ($\hat{t}$) and adopts any strategy ($r(c)$) for reporting realized costs.[17] The cost revelation constraints (3) ensure that the producer will truthfully report realized operating costs, whatever the previous bidding behavior.[18]

II. Properties of the Optimal Bidding Scheme

In this section, we characterize the optimal bidding scheme. Our main result is a theorem that presents analytic expressions for the optimal price, production subsidy, and franchise fee. Corollaries are also stated to emphasize the important qualitative prop-

[17]Our concern throughout is with the Nash equilibrium in which all firms truthfully announce their valuations of the franchise. We do not address the interesting issue of multiple equilibria. (Some thoughts on the issue in a related setting can be found in Joel Demski and Sappington, 1984.)

[18]Note that by constraint (3), $r(c)=c$ is a dominant strategy for the producer.

erties of the optimal bidding scheme. Detailed explanations for these results are in Section III.

In characterizing the solution to the regulator's problem, it is helpful to identify the producer's profit calculations. Recall that monetary compensation for the producer is from two sources. First, the revenue generated from selling the product at the allowed unit price, $p(c,t)$, is retained by the producer. Second, the producer receives a production subsidy, $S(c,t)$, from the regulator. Notice that both the price and production subsidy can depend on the producer's bid as well as his subsequent cost report. Charges against the firm's profit are also of two sources. First, production costs, $cQ(c,t) + K$, are incurred. Second, the producer pays a franchise fee, $\phi(t)$, to the regulator. Notice that the franchise fee depends only on the producer's bid.

Before we can present a complete characterization of the solution to [RP], a regularity condition is needed to ensure that the necessary conditions for a solution to [RP] are also sufficient. Thus, condition (RC) is stated in terms of adjusted marginal cost.

DEFINITION: *Adjusted marginal cost, $m(c,t)$, is given by*

$$m(c,t) = c + [1-t]\frac{F_2(c|t)}{F_1(c|t)}.$$

(RC). *$m(c,t)$ is monotonically increasing in c and monotonically decreasing in t.*[19]

When this regularity condition holds, the important properties of the solution to the regulator's problem are readily characterized, as reported in the following theorem.

THEOREM 1: *If* (RC) *holds, the solution to* [RP] *has the following properties* $\forall c \in [\underline{c}, \bar{c}]$ *and* $\forall t \in [0,1]$:

$$(4) \quad p(c,t) = m(c,t);$$

$$(5) \quad S(c,t) = -\{[p(c,t)-c]\,Q(c,t)\} + \int_c^{\bar{c}} Q(b,t)\,db;$$

$$(6) \quad \phi(t) = \int_{\underline{c}}^{\bar{c}} Q(b,t)F(b|t)\,db - \int_0^t \int_{\underline{c}}^{\bar{c}} \left[\frac{s}{t}\right]^{n-1} Q(b,s)F_2(b|s)\,db\,ds - K.$$

One of the main implications of Theorem 1 is an important separation property: competition influences the optimal scheme only through the franchise fee. That is, for any cost realization and bid, the price and production subsidy are independent of the number of bidders.[20] In particular, the producer faces the same contract he would face in the absence of competition. With competition, he simply pays more up front for this contract. These observations are recorded as Corollaries 1 and 2.

COROLLARY 1: *If* (RC) *holds, the prices $p(c,t)$ and production levels $Q(c,t)$ that constitute the solution to* [RP] *are independent of the number of bidders ($n \geq 1$). The production subsidy, $S(c,t)$, is also independent of the number of bidders.*[21]

COROLLARY 2: *In the solution to* [RP], *the franchise fee, $\phi(t)$, is a nondecreasing function of the number of bidders, n. Consequently, for any winning bid, t, the producer's expected profit at the time the franchise is*

[19]Regularity condition (RC) will be satisfied in a variety of settings. For example, suppose $F(c|t) = tF^A(c) + [1-t]F^B(c)$ where $F^A(c) \geq F^B(c)\ \forall c \in [0,1]$, and let $F_1^i(c) = I^i - \alpha^i c$ for $i = A, B$, where $I^A > I^B$, $1 > \alpha^A > \alpha^B > 0$, and, so that $F_1^A(c)$ and $F_1^B(c)$ are density functions, $I^i = 1 + \frac{1}{2}\alpha^i$, $i = A, B$. Then it is straightforward to verify that regularity condition (RC) is satisfied.

[20]An analogous conclusion is reported by Laffont and Tirole (1985) and McAfee and McMillan (1987a).

[21]Because our conclusions hold for the particular case of $n = 1$, the analysis generalizes Baron and Myerson's model of regulation to a setting where the regulated firm has superior but imperfect information concerning its production technology at the outset of its interaction with the regulator.

awarded is a nonincreasing function of the number of bidders.[22]

Next observe from condition (4) that the markup of price over marginal cost depends on both the *ex post* cost realization, c and the winning bid, t. Price is equal to marginal cost only for the bidder with the highest possible valuation, or for *ex post* cost realizations such that $F_2(c|t)=0$, including extreme cost realizations. These facts are recorded in Corollary 3.

COROLLARY 3: *If* (RC) *holds, price is established at the level of realized marginal cost in the solution to* [RP] *when* (i) $t=1$; (ii) $c=\underline{c}$; (iii) $c=\bar{c}$; *and* (iv) $F_2(c|t)=0$. *Otherwise, price is set above marginal cost, so that realized production levels fall short of levels that are ex post efficient.*

Next consider the producer's actual rents. It is apparent that realized profit gross of the franchise fee is given by the last term on the right-hand side of condition (5) of Theorem 1. Note that the magnitude of this expression increases as c decreases. Thus, as might be expected, since realized costs are privately observed by the producer, his profit must be greater the smaller are realized production costs. If this were not the case, the producer could always exaggerate costs with impunity. These observations are recorded in Corollary 4.

COROLLARY 4: *In any feasible solution to* [RP], *the realized profit for the producer is equal to* $\int_c^{\bar{c}} Q(b,t)\,db-\phi(t)-K$, *which is a nonincreasing function of c for any bid t.*

The producer's *expected* profit is now readily calculated by taking expectations of the expression in Corollary 4. Doing so, and observing condition (6) of Theorem 1, leads immediately to Corollary 5.

COROLLARY 5: *In any feasible solution to* [RP], *the producer's expected profit at the time the franchise is awarded is given by*

$$\int_0^t \int_{\underline{c}}^{\bar{c}} [s/t]^{n-1} Q(b,s) F_2(b|s)\,db\,ds$$

for all t. This expression is nonincreasing in the number of bidders, for any winning bid t.[23]

Additional insight concerning the optimal franchise fee is gained by manipulating condition (6) of Theorem 1. As is evident from Corollary 6, the fee can be decomposed alternatively into two parts. The first part is the producer's estimate of the value of the franchise (gross of the franchise fee) to the next highest bidder. The second part represents the gains from sorting which arise because the allowed price and production subsidy (and, therefore, quantity) vary with the winning bid (i.e., because $Q_2(c,t)\neq 0$).[24]

COROLLARY 6: *In any feasible solution to* [RP], *the franchise fee is also given by*

$$\phi(t)=\left\{\int_0^t \int_{\underline{c}}^{\bar{c}} \left[\frac{n-1}{t}\right]\left[\frac{s}{t}\right]^{n-2} Q(b,s) \times F(b|s)\,db\,ds-K\right\} + \left\{\int_0^t \int_{\underline{c}}^{\bar{c}} \left[\frac{s}{t}\right]^{n-1} Q_2(b,s) F(b|s)\,db\,ds\right\}.$$

[22] The proof follows immediately from condition (6) of Theorem 1, from Corollary 1, and the facts that $Q(c,t)\geq 0$, and $F_2(c|t)\geq 0$ $\forall c\in[\underline{c},\bar{c}]$ and $\forall t\in[0,1]$.

[23] It also follows that expected profits before bidding, $t^{n-1}\Pi(t)$, must be nondecreasing in t.

[24] By Corollary 4, the expected value of the franchise (gross of the franchise fee and the fixed cost of production) for a bidder with valuation s is

$$\int_{\underline{c}}^{\bar{c}} \int_c^{\bar{c}} Q(b,s) F_1(c|s)\,db\,dc=\int_{\underline{c}}^{\bar{c}} Q(b,s) F(b|s)\,db.$$

$[n-1]s^{n-2}/t^{n-1}$ is the conditional density for the second-highest-order statistic (s) given that the maximal order statistic is t. See also fn. 27.

Thus, in general, the linkage of the bidding and production stages allows the regulator to secure a greater up-front payment from the producer than would be possible under the L-M franchise bidding scheme. However, there are conditions under which this simple and attractive scheme is indeed the optimal one. These conditions are summarized in Corollary 7.

COROLLARY 7: *If* $F_2(c|t)=0\ \forall c\in[\underline{c},\bar{c}]$ *and* $t\in[0,1]$, *and/or if* $n=\infty$, *then in the solution to* [RP]: (i) *expected profits of all bidders are zero, and* (ii) *the output level of the producer is ex post efficient for all cost realizations.*

Thus, the L-M scheme constitutes the solution to [RP] in the limiting cases where 1) the private signals of bidders are uninformative, so that all bidders are identical, and/or 2) where there are an infinite number of bidders, so that the winning bidder is certain to have the highest possible valuation, $t=1$. More generally, though, optimal prices are above realized marginal cost and depend not only on realized costs but also on initial cost projections (i.e., bids).

III. Explanations

There are a number of major conclusions reported in the preceding section. First, the regulator benefits from linking the bidding and production stages, so that production and compensation schedules vary with the winning bid. Second, regulated prices are set equal to adjusted marginal cost, which generally exceeds realized marginal cost; hence, production distortions are introduced. Third, for any $\{c,t\}$ realization, these distortions are independent of the number of bidders. The explanations for these findings lie in the fact that there are two distinct incentive problems that concern the regulator. First, he must dissuade bidders from understating their true valuations of the franchise (i.e., exaggerating expected costs via understating t). Second, he must subsequently prevent the producer from exaggerating realized production costs. The first of these problems dictates the magnitude of the franchise fee. The second determines the production subsidy and allowed price.[25]

As Corollary 4 reports, the producer's realized profit must be greater the smaller are realized production costs, whatever his bid may have been. Thus, the producer can only be induced to truthfully report realized costs with a production subsidy that makes realized profits, $\int_c^{\bar{c}}Q(b,t)\,db-\phi(t)-K$, greater the smaller are reported costs.[26] The key point to note is that realized profits rise more rapidly as c falls the greater are induced production levels, $Q(b,t)$. The reason is that larger production subsidies provide the incentive for higher levels of output.

A critical question from the regulator's perspective is whether anticipated rents from the production subsidy can be recaptured entirely through the franchise fee. This would be the case if the bidders' valuations were known to the regulator. Similarly, if potential producers were identical *ex ante*, they would bid the entire expected value of the production subsidies in an attempt to obtain the monopoly franchise. Consequently, the L-M scheme is optimal in this setting. Furthermore, with n large, many bidders are likely to have very similar valuations of production subsidies, so the L-M scheme again performs well. This is the essence of Corollary 7.

More generally, however, with a finite number of diverse bidders, all rents cannot be recovered through the franchise fee. In equilibrium, under a sealed-bid, first-price auction, each potential producer bids his

[25] Both of these problems might be readily resolved if a public, informative monitor of realized production costs were available. See our paper (1986a) for the details of how the optimal incentive scheme would be constructed when the firms are of a finite number of types. Also see Baron and David Besanko (1984b). Riley (1985) also considers a bidding model with public *ex post* information. He shows that bidding over royalty rates and fee bidding with preset royalties can increase the expected revenue of the auctioneer relative to bidding only over fees. Also see Douglas Reece (1979).

[26] The optimal subsidy schedule is identical to that in Baron and Myerson. They interpret the production subsidy as the consumers' surplus associated with an "adjusted demand function." Thus, the optimal scheme could be interpreted as an "adjusted" L-M scheme.

assessment of the expected valuation of the next most efficient firm. Thus, if the L-M scheme were adopted, the regulator would forego rents equal to the expected difference between the highest and second-highest valuations.[27]

The essence of our findings is that rents are limited more successfully by departing from the L-M scheme, and reducing production levels. The manner in which output distortions limit rents is subtle. Consider the L-M scheme as a benchmark. Then suppose that for a particular bid t, the regulator reduced production subsidies for low-cost realizations and simultaneously reduced the franchise fee so as to leave unchanged the expected profit of a firm with valuation t. These compensating changes reduce the attraction to more efficient firms of misrepresenting their valuation as t. The reason is that lower-cost realizations are more likely for more efficient firms, so these firms are more likely to be adversely affected by the subsidy reductions. Consequently, the franchise fees for more efficient firms can be raised without having them misrepresent their valuations as t. And though the reduced production subsidies reduce production levels when t is the winning bid, the resulting direct loss in consumers' surplus is of second-order importance for small variations, while the gain from higher franchise fees is of first-order importance. Thus departures from L-M scheme are welfare enhancing. The gains from higher franchise fees that result from production distortions are reflected in Corollary 6.

The pricing equation (4) defines how distortions are optimally employed to limit rents. The equation can be written as

$$(4') \quad [W'(Q(c,t))-c]\,nt^{n-1}F_1(c|t) = nt^{n-1}[1-t]F_2(c|t).$$

The left-hand side is the incremental gain in social surplus from an output expansion, weighted by the probability that the gain will actually be realized. The right-hand side is the corresponding reduction in expected franchise fees.[28] Hence, the optimal distortions are those which equate the marginal benefits and marginal costs of output expansion.

Output distortions are most effective as a sorting device where their *differential* impact on the expected profits of bidders is greatest. Thus, in regions where the probability of achieving a cost realization below c rises very rapidly with t (i.e., where $F_2(c|t)$ is large), relatively large distortions in $Q(c,t)$ are desirable, as they constitute a strong deterrent to underbidding. On the other hand, in regions where $F_2(c|t)$ is small, the impact of quantity distortions on the expected profit of bidders does not vary much according to their true valuation, and the optimal distortions are therefore relatively small. In the limiting case where all bidders value the franchise identically (i.e., where $F_2(c|t)=0\ \forall c,t$), no quantity distortions are implemented.

Thus, as is evident from property (4) of Theorem 1, markups of price $p(c,t)$ above marginal cost c are proportional to $F_2(c|t)$. The markups vary inversely with the probability that the particular $\{c,t\}$ pair will actually be realized, as this is the likelihood that the direct cost of the induced output distortion will actually be incurred. The magnitude of the induced distortion is also influenced by the expected increase in franchise fees that might be collected from a producer with a valuation exceeding t.

[27]Thus, if production levels did not depend on t, i.e., $Q(c,t)=\tilde{Q}(c)$ (as in the L-M scheme), the franchise fee would correspond to the equilibrium bid function of an ordinary first-price sealed-bid auction, in which an object that has value $v(t)=\int_{\underline{c}}^{\bar{c}}\tilde{Q}(c)F(c|t)\,dc$ to a bidder of type t is auctioned off. See McAfee and McMillan (1987b).

[28]From property (6) in Theorem 1, $|\partial\phi(s)/\partial Q(c,t)| = [t/s]^{n-1}F_2(c|t)$ represents the change in the franchise fee that can be charged the bidder with valuation s ($s>t$), without having him misrepresent his valuation as t, when $Q(c,t)$ is raised incrementally. Thus, the total expected change in franchise fees from raising $Q(c,t)$ is

$$\int_t^1\left|\frac{\partial\phi(s)}{\partial Q(c,t)}\right|ns^{n-1}\,ds = nt^{n-1}[1-t]F_2(c|t).$$

Hence, the $F_2(\cdot)/F_1(\cdot)$ ratio in the optimal markup of price above marginal cost is discounted by the term $[1-t]$, which reflects the range of possible bidders with higher valuations.[29] Thus, as the number of bidders becomes infinitely large, the probability that the highest valuation is the most favorable approaches unity, and no production distortions are induced. These observations explain Corollary 3.

Finally, recall that this optimal markup and the corresponding production subsidies are independent of the number of bidders (Corollary 1). This result obtains because increased competition has two offsetting effects. First, with a greater number of competitors, each bidder is less likely to shade his bid, since such shading runs a greater risk of losing the profitable franchise to a competitor. Thus, competitive pressure can substitute for production distortions that would otherwise be employed to elicit truthful bids. Second, a given production distortion in $Q(c,t)$ $(t<1)$ has a greater expected impact because it is more likely that bidders will have valuations higher than t, and will therefore be affected by the distortions. This second effect, in isolation, would result in more pronounced production distortions as n increased. But when bidders have independent private valuations of the franchise, the second effect of increased competition exactly offsets the first, and production distortions for any $\{c,t\}$ realization are independent of the number of bidders. Competition certainly benefits the regulator, but the gains from competition arise solely in the form of higher franchise fees (Corollary 2).

[29]Baron and Besanko offer a related interpretation of the ratio $F_2(c|t)/F_1(c|t)$. They refer to this fraction as a local measure of "informativeness," and note that it measures by "how much $[c]$ must be increased when $[t]$ is [reduced] to maintain the same probability that marginal cost will be no greater than the resulting $[c]$" (1984a, p. 286). In effect, informativeness refers to the extent of cost exaggeration necessary to make the report of c "consistent" with an earlier understatement of t. This interpretation accents the parallels between the notion of informativeness and Laffont and Tirole's 1986 notion of a concealment set.

IV. Summary and Conclusions

We have derived the optimal procedure for awarding a monopoly franchise in a setting where risk-neutral firms initially have imperfect private knowledge of prospective production costs. The optimal procedure was referred to as "menu contracting," as the regulator optimally offers a menu of production contracts to potential producers. The contract that is ultimately implemented depends on the winning bid. A higher winning bid elicits a contract that requires the producer to pay a larger franchise fee up front, but provides more generous production subsidies. Tailoring the production contract to the winning bid discourages firms from shading their bids, since winning the franchise with a lower bid gains a production contract that is less profitable.

In closing, we comment briefly on some potentially interesting extensions of our analysis. As noted in the introduction, our "one-shot" model does not account for contracting costs; nor does it allow for unforeseen contingencies. Furthermore, though the regulator's information is imperfect in our model, it is really quite detailed. A dynamic model that incorporated a less omniscient regulator (and perhaps one with less commitment ability) likely would provide policy prescriptions that differ from ours in detail, but perhaps not in spirit.

Also recall that the private signals of bidders were assumed to be uncorrelated. With correlated information, it would generally be optimal for the regulator to base the payment to each bidder on all announced bids. The work of Jacques Cremèr and Richard McLean (1985a,b) suggests that in this setting, the regulator might generically be able to extract all surplus from bidders and effect efficient production levels.[30] In theory, this could be done by treating each firm's bid as

[30]Their formal analysis, which presumes the absence of bankruptcy concerns, also presumes a finite number of possible cost realizations and private signals. Whether their insight generalizes to the case of continuous distributions remains an open question.

a prediction about other bids, and penalizing bidders for incorrect predictions.

The penalties involved in this bidding procedure can be immense, however, and where bankruptcy laws prevent enforcement of excessive penalties, the regulator will not be able to effect his most preferred outcome. Inability to do so will generally result in distortions from efficient production levels over ranges where bankruptcy constraints are binding, much as in the case of independent private signals considered in this paper. The distortions mitigate incentives to misrepresent cost realizations and valuations of the franchise whenever it is too costly to eliminate such misrepresentation through payments (and penalties) alone.[31,32]

Finally, we note that other forms of distortions may arise under different assumptions about the environment. For example, despite increasing returns to scale, the regulator might find it optimal to award the franchise to two or more firms simultaneously if sunk costs are sufficiently small.[33] Doing so can foster competition *ex post* and thereby lessen the rents the producer might otherwise command from private cost information. As another example, it is conceivable that the regulator might not always award the franchise to the firm that is most efficient, *ex ante*. The work of Myerson (1981) and McAfee and McMillan (1985) indicates that such a distortion may be optimal if bidders differ over observable characteristics.

[31] There are other reasons why, in practice, the payment to a bidder will not depend on the announcements of other bidders. Such a payment structure could afford the auctioneer the opportunity to employ a shill to strategically place bids so as to unfairly extract rents from the bidders. Similarly, there are reasons why, in practice, losing bidders will not be penalized. In cases where the auctioneer has superior information about the value of the object being auctioned, the possibility arises that the primary purpose of the auction could be to collect entry fees from losing bidders.

[32] For details of an analysis that explicitly accounts for bankruptcy constraints, see our 1986b paper.

[33] This possibility is noted in Barry Nalebuff and Stiglitz (1983). See also Anton and Yao.

Appendix: Proof of Theorem 1

For expositional simplicity, we assume $m(c,t)$ is differentiable in both its arguments, which ensures the optimal $Q(c,t)$ and $T(c,t)$ are similarly differentiable. A more general proof along the lines of Myerson (1981) is available from the authors. To begin, define

$$\text{(A1)} \quad \tilde{\Pi}(b,c|t) = [P(Q(b,t)-c] \times Q(b,t) + T(b,t) - K.$$

The cost revelation constraints (3) require $c = \operatorname{argmax}_b \tilde{\Pi}(b,c|t)$, for all c, t. Thus, letting $\Pi(c|t) \equiv \tilde{\Pi}(c,c|t)$, the envelope theorem provides

$$\text{(A2)} \quad \partial \Pi(c|t)/\partial c = -Q(c,t).$$

Integration then allows us to write $\Pi(c|t)$ as

$$\text{(A3)} \quad \Pi(c|t) = A(t) + \int_c^{\bar{c}} Q(b,t)\,db,$$

for some function $A(t)$.

Note that $\partial Q(c,t)/\partial c \le 0$ combined with (A2) is necessary and sufficient to ensure truthful reporting of c.

Next, define

$$\text{(A4)} \quad \tilde{V}(s|t) \equiv [s]^{n-1} \int_{\underline{c}}^{\bar{c}} \Pi(c|s) F_1(c|t)\,dc.$$

From (A2), we have

$$\text{(A5)} \quad \tilde{V}(s|t) = [s]^{n-1} \times \left\{ \int_{\underline{c}}^{\bar{c}} Q(c,s) F(c|t)\,dc + \Pi(\bar{c}|s) \right\}.$$

The truthful bidding constraints (2) require $t = \operatorname{argmax}_s \tilde{V}(s|t)$, $\forall\ t$. Hence, letting $V(t) \equiv \tilde{V}(t|t)$, the envelope theorem provides

$$\text{(A6)} \quad V'(t) = [t]^{n-1} \int_{\underline{c}}^{\bar{c}} Q(c,t) F_2(c|t)\,dc.$$

It follows from (A6) that $\partial Q(c,t)/\partial t \ge 0$ is sufficient to ensure truthful bidding globally.

Since $V'(t) \geq 0$, the individual rationality constraints (1) imply $V(0)=0$. Therefore, integrating (A6) yields

$$\text{(A7)} \quad V(t)=\int_0^t \int_{\underline{c}}^{\bar{c}} s^{n-1} Q(c,s) F_2(c|s)\, dc\, ds.$$

Now, combining (A3), (A4), and (A7), and letting $\phi(t) \equiv K - A(t)$ provides

$$\text{(A8)} \quad \phi(t)=\int_{\underline{c}}^{\bar{c}} Q(c,t) F(c|t)\, dc - \int_0^t \int_{\underline{c}}^{\bar{c}} \left[\frac{s}{t}\right]^{n-1} Q(b,s) F_2(b|s)\, db\, ds - K,$$

which is condition (6) of Theorem 1. Condition (5) then follows from (A1) and (A3).

Finally, consider the regulator's problem. The regulator's objective is to maximize

$$\begin{aligned} B &\equiv \int_0^1 \int_{\underline{c}}^{\bar{c}} [W(Q(c,t)) - T(c,t)]\, F_1(c|t) n t^{n-1}\, dc\, dt \\ &= n\int_0^1 \left\{ \int_{\underline{c}}^{\bar{c}} [W(Q(c,t)) + P(Q(c,t)) \times Q(c,t) - c\, Q(c,t)]\, F_1(c|t) - [1-t] Q(c,t) F_2(c|t)\, dc \right\} t^{n-1}\, dt. \end{aligned}$$

The second equality follows from (A1) and (A7). Then, using the definition of consumers' surplus, pointwise maximization yields the necessary conditions:

$$\text{(A9)} \qquad P(Q(c,t)) - m(c,t) = 0.$$

Condition (4) of Theorem 1 follows from $P(Q(c,t)) \equiv p(c,t)$. (RC) guarantees $\partial Q(c,t)/\partial c \leq 0$ and $\partial Q(c,t)/\partial t \geq 0$.

REFERENCES

Anton, James and Yao, Dennis, "Second Sourcing and the Experience Curve: Price Competition in Defense Procurement," *Rand Journal of Economics*, Spring 1987, forthcoming.

Baron, David and Besanko, David, (1984a) "Regulation and Information in a Continuing Relationship," *Information Economics and Policy*, June 1984, *1*, 267–302.

______ **and** ______, (1984b) "Regulation, Asymmetric Information, and Auditing," *Rand Journal of Economics*, Winter 1984, *15*, 447–70.

______ **and Myerson, Roger,** "Regulating a Monopolist with Unknown Costs," *Econometrica*, July 1982, *50*, 911–30.

Caillaud, B. et al., "The Normative Economics of Government Intervention in Production in the Light of Incentives Theory: A Review of Recent Contributions," mimeo., MIT, March 1985.

Cremèr, Jacques and McLean, Richard, (1985a) "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," mimeo., University of Pennsylvania, June 1985.

______ **and** ______, (1985b) "Optimal Selling Strategies Under Uncertainty for a Discriminating Monopolist when Demands are Interdependent," *Econometrica*, March 1985, *53*, 345–62.

Dasgupta, Partha, Hammond, Peter and Maskin, Eric, "The Implementation of Social Choice Rules: Some Results on Incentive Compatibility," *Review of Economic Studies*, April 1970, *46*, 185–216.

Demsetz, Harold, "Why Regulate Utilities?," *Journal of Law and Economics*, February 1968, *11*, 55–65.

Demski, Joel and Sappington, David, "Optimal Incentive Contracts with Multiple Agents," *Journal of Economic Theory*, June 1984, *33*, 152–71.

Laffont, Jean-Jacques and Tirole, Jean, "Auctioning Incentive Contracts," mimeo., MIT, October 1985.

______ **and** ______, "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, June 1986, *94*, 614–41.

Loeb, Martin and Magat, Wesley, "A Decentralized Method for Utility Regulation," *Journal of Law and Economics*, August 1979, *22*, 399–404.

McAfee, R. Preston and McMillan, John, "Discrimination in Auctions," mimeo., Univer-

sity of California-San Diego, January 1985.

______ **and** ______, "Bidding for Contracts: A Principal-Agent Analysis," *Rand Journal of Economics*, Autumn 1986, *17*, 326–38.

______ **and** ______, (1987a) "Competition for Agency Contracts," *Rand Journal of Economics*, forthcoming 1987.

______ **and** ______, (1987b) "Auctions and Bidding," *Journal of Economic Literature*, forthcoming 1987.

Milgrom, Paul, "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, Autumn 1981, *12*, 380–91.

______ **and Weber, Robert,** "A Theory of Auctions and Competitive Bidding," *Econometrica*, September 1982, *50*, 1089–122.

Myerson, Roger, "Incentive Compatibility and the Bargaining Problem," *Econometrica*, January 1979, *47*, 61–74.

______, "Optimal Auction Design," *Mathematics of Operations Research*, February 1981, *6*, 58–73.

Nalebuff, Barry and Stiglitz, Joseph, "Information, Competition and Markets," *American Economic Review Proceedings*, May 1983, *73*, 278–83.

Reece, Douglas, "An Analysis of Alternative Bidding Schemes for Leasing Offshore Oil," *Bell Journal of Economics*, Autumn 1979, *10*, 659–69.

Riley, John, "Ex Post Information and Auctions," mimeo., UCLA, August 1985.

______ **and Samuelson, William,** "Optimal Auctions," *American Economic Review*, June 1981, *71*, 381–92.

Riordan, Michael, "On Delegating Price Authority to a Regulated Firm," *Rand Journal of Economics*, Spring 1984, *15*, 108–15.

______ **and Sappington, David,** (1986a) "Optimal Contracts with Public and Private Ex Post Information," Stanford University and Bell Communications Research mimeo., rev. January 1986.

______ **and** ______, (1986b) "Designing Procurement Contracts," mimeo., Stanford University and Bell Communications Research, May 1986.

Sappington, David and Stiglitz, Joseph, "Information and Regulation," in E. Bailey, ed., *Public Regulation: New Perspectives on Institutions and Policies*, Cambridge: MIT Press, 1987.

Spulber, Daniel, "Bargaining and Regulation with Asymmetric Information about Demand and Supply," *Journal of Economic Theory*, forthcoming 1987.

Williamson, Oliver, "Franchise Bidding for Natural Monopolies—In General and with Respect to CATV," *Bell Journal of Economics*, Spring 1976, *7*, 73–104.

______, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

Contracts as a Barrier to Entry

By PHILIPPE AGHION AND PATRICK BOLTON*

It is shown that an incumbent seller who faces a threat of entry into his or her market will sign long-term contracts that prevent the entry of some lower-cost producers even though they do not preclude entry completely. Moreover, when a seller possesses superior information about the likelihood of entry, it is shown that the length of the contract may act as a signal of the true probability of entry.

Most of the literature on entry prevention deals with the case of two duopolists (the established firm and the potential entrant) who compete with each other to share a market, where one of the duopolists (the incumbent) has a first-move advantage.[1] This basic paradigm has been studied under various assumptions: about the strategy space of the players; the information structure of the game; and the time horizon. Recently, the model has been enlarged to allow for several entrants, several incumbents, several markets, and third parties.[2]

We propose here to extend the entry-prevention model in one other direction, which to our knowledge has not yet been formalized; namely, we consider whether optimal contracts between buyers and sellers deter entry and whether they are suboptimal from a welfare point of view. It has been pointed out by many economists that contracts between buyers and sellers in intermediate-good industries may have significant entry-prevention effects and that such contracts may be bad from a welfare point of view.[3]

On the other hand, it is a widespread opinion among antitrust practitioners that contracts between buyers and sellers are socially efficient.[4] There have been a number of antitrust cases involving exclusive dealing contracts and often the decision reached by the judge has lead to considerable controversy. One famous case, *United States v. United Shoe Machinery Corporation* (1922), illustrates quite clearly the nature of the debate: the United Shoe Machinery Corporation controlled 85 percent of the shoe-machinery market and had developed a complex leasing system of its machines to shoe manufacturers, a leasing system against which, it was thought, other machinery manufacturers would have difficulty competing. The judge ruled that these leasing contracts were in violation of the Sherman Act; his decision has been repeatedly criticized by leading antitrust experts (see Richard Posner, 1976, and Robert Bork, 1978). The main argument against the decision has been expressed by Posner: "The point I particularly want to emphasize is that the customers of United would be unlikely to participate in a campaign to strengthen United's monopoly position without insisting on being compensated for the loss of alternative and less costly

*Department of Economics, Harvard University, Cambridge, MA 02138, and University of California, Berkeley, CA 94720, respectively. We are greatly indebted to Jean Tirole for helping us formulate the model. We also thank Jerry Green and Oliver Hart for their useful suggestions and kind encouragement. We have been most fortunate to benefit from many helpful discussions with Dilip Abreu, Richard Caves, Nancy Gallini, Andreu Mas-Colell, and Eric Maskin.

[1]See, for example, the seminal contributions by Michael Spence (1977) and Avinash Dixit (1979, 1980).

[2]For a recent survey, see Drew Fudenberg and Jean Tirole (1986).

[3]Spence (p. 544), for example, briefly mentioned contracts as a method for impeding entry; see also Oliver Williamson (1979). Furthermore, there is a literature on barriers to entry and vertical integration that is relevant to our discussion, since most of the time what vertical integration achieves in this literature can also be done through an appropriate contract. (See Roger Blair and David Kaserman, 1983.)

[4]This position has been forcefully defended by Robert Bork (1978), for example.

(because competitive) sources of supply" (p. 203). Exactly the same point is made by Bork (p. 140), who concludes that when we find exclusive dealing contracts in practice, then these contracts could not have been signed for entry-deterrence reasons.

Both Posner and Bork are right in pointing out that the buyer is better off when there is entry and that he (she) will tend to reject exclusive dealing contracts that reduce the likelihood of entry unless the seller compensates him (her) by offering an advantageous deal. Nevertheless, we show that contracts between buyers and sellers will be signed for entry-prevention purposes.

When the buyer and the seller sign a contract, they have a monopoly power over the entrant. They can jointly determine what fee the entrant must pay in order to be able to trade with the buyer; that is to say, if the buyer signs an exclusive contract with the seller and then trades with the entrant, he must pay damages to the seller. Thus he will only trade with the entrant if the latter charges a price which is lower than the seller's price minus the damages he pays to the seller. These damages, which are determined in the original contract (liquidated damages), act as an entry fee the entrant must pay to the seller. We show that the buyer and the seller set this entry fee in the same way that a monopoly would set its price, when it cannot observe the willingness to pay of its customers. Thus, the main reason for signing exclusive contracts, in our model, is to extract some of the surplus an entrant would get if he entered the seller's market.

These contracts introduce a social cost, for they sometimes block the entry of firms that may be more efficient than the incumbent seller. Entry is blocked because the contract imposes an entry cost on potential competitors. This cost takes two different forms: an entrant must either wait until contracts expire, or induce the customers to break their contract with the incumbent by paying their liquidated damages.

The waiting cost is larger, other things being equal, the longer the contract. We are thus led to study the question of the optimal length of the contract. It is a well-known principle in economics that if agents engage in mutually advantageous trade, it is in their best interest to sign the longest possible contract. A long-term contract can always replicate what a sequence of short-term contracts achieves.

This principle, however, sharply contrasts empirical evidence: In practice most contracts are of an explicit finite duration. Many economists have been puzzled by this obvious discrepancy between the theory and empirical evidence, and several authors have attempted to provide an explanation for why contracts are of a finite duration; most notably Oliver Williamson (1975, 1979) and Milton Harris and Bengt Holmström (1983).

We argue here that looking only at the length of a contract is misleading. What is important is to what extent a contract of a given length locks the parties into a relationship. Thus we are led to make the distinction between the *nominal length* of the contract (the length that is specified in the contract) and the *effective length* of the contract (the actual length that the parties expect the relationship to last at the time of signing). Liquidated damages constitute an implicit measure of the effective length of the contract.

The paper is organized as follows: Section I looks at optimal contracts between a single buyer and the incumbent seller, when both parties have the same information about the likelihood of entry. Section II analyzes optimal contracts when there is asymmetric information about the probability of entry. Section III deals with optimal contracts when there are several buyers. Finally, Section IV offers some concluding comments.

I. Optimal Contracts Between One Buyer and the Incumbent Seller

We consider a two-period model, where a single producer supplies one unit to a buyer. The latter has a reservation price, $P=1$, and buys at most one unit. The seller faces a threat of entry, which is modeled as follows: At the time of contracting the seller's unit cost is $c=\frac{1}{2}$, while the entrant's cost of producing the same homogenous good is not known. For simplicity we assume that the entrant's cost, c_e, is uniformly distributed in

$[0,1]$.[5] Furthermore, if entry occurs and no contract has been signed between the incumbent and the buyer, both suppliers compete in prices, so that the Bertrand equilibrium price is given by $P = \max\{\frac{1}{2}, c_e\}$. When there is no entry, the potential entrant makes zero profits. Thus entry will only occur if $c_e \leq \frac{1}{2}$ and the probability of entry is given by

$$(1) \qquad \phi = Pr\left(c_e \leq \tfrac{1}{2}\right) = \tfrac{1}{2}.$$

We attempt here to model in the simplest way the view of the world where there are many investors at each period of time who try to invest their funds in the markets where they hope to get the highest returns. The distribution of profits across markets, however, changes stochastically over time. Therefore entry into a given market may also be stochastic. In this story it is implicitly assumed that investors do not have an unlimited access to funds and/or that there are diminishing returns to managing more investment projects. If neither of these assumptions hold, then investment will take place until the marginal return on the last investment project is equal to the interest rate. Many good reasons have been given for why investors only have a limited access to funds (see for example, Joseph Stiglitz and Andrew Weiss, 1981, or Williamson, 1971).

The timing of the game is as follows: At date 1 the incumbent seller and the buyer negotiate a contract, then entry either takes place or does not. Finally at date 2, there is production and trade.[6] We assume that the entrant's cost, c_e, is *not observable* but the parties to the contract know the distribution function of c_e. Therefore, contracts contingent on c_e cannot be written.[7]

If no contract is signed at date 1, the buyer's expected payoff is given by

$$(2) \qquad (1-\phi)\cdot 0 + \phi\cdot\tfrac{1}{2} = \tfrac{1}{2}\cdot\tfrac{1}{2} = \tfrac{1}{4}.$$

That is, with probability $(1-\phi)$ there is no entry and the seller sets the price equal to one. Hence, the buyer gets no surplus. With probability ϕ, entry occurs and Bertrand competition drives the price down to the incumbent's unit cost $c = \frac{1}{2}$. Now, Posner's point simply was that any contract that is acceptable to the buyer must give him an expected surplus of at least $\frac{1}{4}$ (assuming that the buyer is risk neutral). We shall show that even though the seller faces this constraint, there are gains to signing long-term contracts and in preventing entry.

The buyer and the incumbent seller could conceivably sign very complicated contracts even in this simple setting. For example, the price specified in the contract may be contingent on the event of entry or even contingent on the entrant's offer.[8] We shall, however, restrict ourselves to simple contracts of the form $c = \{P, P_0\}$ and show that there is no loss of generality in considering only this type of contract. Here P is the price of the good when the buyer trades with the incumbent and P_0 is the price the buyer must

[5]The choice of a uniform distribution is entirely for the sake of computational simplicity. In our 1985 paper, we show that the qualitative results obtained here are valid for any continuous density $f(x)$ with a support such that the lower bound is finite and that contains the interval $[0\ \frac{1}{2}]$.

[6]When production takes place before entry, the analysis is slightly modified. When the buyer switches to the entrant, the incumbent must now incur a loss of $c = \frac{1}{2}$. Thus the Bertrand equilibrium in the post-entry game now is $P = c_e$, so that entry will be precluded (since the entrant always makes nonpositive profits). To avoid an outcome where *ex post* competition (after entry) drives out *ex ante* competition (see Partha Dasgupta and Stiglitz, 1984), we then need to assume that the entrant sometimes makes losses when he does not enter into the incumbent's market. In other words, the entrant sometimes has a negative opportunity cost (see our earlier paper).

[7]In general, what matters is not the actual unit cost of the entrant but his opportunity cost of not entering. If one takes this interpretation, then nonobservability of the entrant's opportunity cost is a mild assumption.

[8]One often observes contracts where a retailer provides a minimum price warranty of the form: "If the buyer is offered a lower price by another retailer for the same good, within t periods, he can then claim back the difference between the high and the low price." These are examples of contracts which are contingent on the entrant's offer. Of course, if such contracts are written then entry is precluded (since the entrant makes zero profits). See our discussion of these contracts in Section IV.

pay if he does not trade with the incumbent. In other words, P_0 represents *liquidated damages*.

When a contract $c = \{P, P_0\}$ is signed, the buyer gets a surplus of $1-P$ if there is no entry. Furthermore, if there is entry, he will only switch to the entrant if the latter offers a surplus of at least $1-P$. We shall assume that when the buyer is indifferent between switching and not switching, he trades with the entrant. Thus in the post-entry equilibrium, the buyer also gets a surplus of $1-P$. Then a contract $c=\{P, P_0\}$ is acceptable to the buyer only if

$$(3) \qquad 1-P \geq \tfrac{1}{4}.$$

Next, an entrant can only attract the buyer if he sets a price $\tilde{P}$, such that

$$(4) \qquad \tilde{P} \leq P - P_0$$

(in equilibrium the entrant sets, $\tilde{P} = P - P_0$). And entry only occurs if the entrant makes positive profits:

$$(5) \qquad \tilde{P} - c_e \geq 0.$$

Thus, when a contract $c = \{P, P_0\}$ is signed the probability of entry becomes

$$(6) \qquad \phi' = \max\{0; P - P_0\}.$$

The incumbent now faces the following program:

$$(7) \qquad \max_{P, P_0} \phi' \cdot P_0 + (1-\phi')(P-c),$$

$$\text{subject to} \qquad 1-P \geq \tfrac{1}{4}.$$

It is straightforward to verify that the optimal contract is then given by $c = \{\frac{3}{4}; \frac{1}{2}\}$.

There are several conclusions to be drawn. First, the incumbent's expected payoff of signing the contract $c = \{\frac{3}{4}, \frac{1}{2}\}$ is given by $\pi = \frac{1}{16} + \frac{1}{4}$. If he had not signed a contract, or if he had signed a contract that completely blocks entry, his expected payoff would be $\frac{1}{4}$. Hence he is strictly better off signing this contract and the buyer is not worse off.

Second, when $c = \{\frac{3}{4}, \frac{1}{2}\}$ is signed, the probability of entry is $\phi' = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$. Thus the optimal contract prevents entry to some extent but does not preclude entry completely. The contract $c = \{P, P_0\}$ changes the entry game in a subtle way. On the one hand, it sets a large entry fee, P_0, to the entrant. This reduces the likelihood of entry. But $P_0 = \frac{1}{2}$, does not completely eliminate entry, since the contract commits the incumbent to set a price $P = \frac{3}{4}$. Thus all entrants with costs $c_e \leq \frac{1}{4}$ will find it profitable to enter. Furthermore, even if the incumbent had the opportunity of lowering the price P below $\frac{3}{4}$ in the post-entry game, he would not want to do this. *The incumbent is strictly better off when the buyer switches to the entrant* in the post-entry game, for then he gets a surplus of $\frac{1}{2}$ compared with a maximum surplus of $P - c = \frac{1}{4}$, if he retained the buyer.

By signing a contract, the incumbent and the buyer form a coalition which acts like a nondiscriminating monopolist with respect to the entrant. The coalition sets P_0 like a monopolist sets its price when it cannot discriminate between buyers with different willingnesses to pay.[9] If c_e were observable, the contract could specify P_0 as a function of c_e and the coalition would be able to extract all of the entrant's surplus ($P_0 = \frac{1}{2} - c_e$).

The idea that the incumbent and the buyer can get together and extract some of the entrant's rent is very general. It does not depend, for instance, on the assumption that the seller sets the contract. Peter Diamond and Eric Maskin (1979) have obtained a similar result in the context of a model of search with breach of contract, where neither the buyer nor the seller has the power of making take-it-or-leave-it offers. Rather, Diamond and Maskin assume that the outcome of the bargaining game between a buyer and a seller is given by the Nash-bargaining solution.

[9]An interesting feature of the optimal contract is that if the probability of entry ϕ increases, then the optimal price P_0 may decrease. For example if the incumbent's unit cost is k then $\phi = k$ and $P_0^* = 1 - k(1-k) - k/2$. Thus $dP_0^*/dk < 0$ for $k < \frac{3}{4}$.

Given that the incumbent and the buyer can only act as nondiscriminating monopolists, with respect to potential entrants, the optimal contract introduces a *social cost*, for it sometimes blocks the entry of a firm with a lower cost of production than the incumbent. When an optimal contract is signed, entrants with costs $c_e \in [\frac{1}{4}; \frac{1}{2}]$ do not enter.

To close this section we explain why the buyer and the seller can restrict themselves to simple contracts, $c = \{P, P_0\}$. The buyer and the seller can form a coalition whose value is $\frac{1}{2}$ when they do not allow entry into the market (the buyer's reservation price is one and the incumbent's cost is $c = \frac{1}{2}$). They can raise their payoff by allowing entry and making the entrant pay a fee, which in general will be a function of the entrant's cost, c_e. But the entrant's cost is private information so that the coalition faces a revelation of information problem. Now, a direct mechanism would specify a transfer from the entrant to the coalition, which is a function of the entrant's cost report: $t(c_e)$. This function $t(c_e)$ must satisfy the incentive-compatibility (IC) constraints: for all $c_e \in [0,1]$,

$$(IC) \quad \pi(c_e) - t(c_e) \geq \pi(c_e) - t(\hat{c}_e) \quad \text{for all } \hat{c}_e \in [0,1].$$

(Where $\pi(c_e)$ is the entrant's rent when his cost is c_e.) The IC constraints imply that $t(c_e) = t$ for all $c_e \in [0,1]$. In other words, the entry fee is independent of the entrant's cost.

Next, the entrant's rent is given by the difference between the incumbent's cost and his cost, c_e (i.e., $\pi(c_e) = \frac{1}{2} - c_e$). The coalition chooses t to maximize:

$$t \cdot Pr(\pi(c_e) \geq t) = t \cdot Pr(\tfrac{1}{2} - c_e \geq t) = t(\tfrac{1}{2} - t).$$

Then the optimal transfer is $t^* = \frac{1}{4}$ and the expected surplus raised is $\frac{1}{16}$. Notice that the optimal contract $c = \{P = \frac{3}{4};\ P_0 = \frac{1}{2}\}$ also raises a surplus of $\frac{1}{16}$ from the entrant. We can now appeal to the revelation principle (Dasgupta, Peter Hammond, and Maskin, 1979), which says that no indirect mechanism does better than the best direct mechanism. That is, no other contract exists that raises a higher surplus than $\frac{1}{16}$. Therefore there is no loss in restricting the contracts to be of the form $c = \{P, P_0\}$.[10]

II. Asymmetric Information About the Probability of Entry

In Section I it was assumed that both the incumbent and the buyer know the true probability of entry. This is not always realistic and one would expect that often the incumbent is better informed about the possibility of entry than the buyer. For example, if the incumbent is a high-tech firm and is the only one to have the know-how to produce a given intermediate good, then it is likely to be much better informed than its customers about the ability of a potential competitor in acquiring this know-how and thus produce the intermediate good. Hence, in this section we assume that the incumbent has some private information about the likelihood of entry.[11]

Asymmetric information has important consequences for the determination of the optimal nominal length of the contract. Under symmetric information, there is no incentive for writing a contract of finite nominal length. On the contrary, the incumbent always gains by locking the buyer into a contract in every period, for then an entrant cannot avoid paying the entry fee by entering at a time when the buyer is not bound by a contract to the incumbent. Under asymmetric information, on the other hand, the seller may wish to sign a contract of finite nominal length in order to signal to the buyer that entry is unlikely. Of course, the seller could also signal his information by

[10] In the above discussion we have restricted ourselves to deterministic mechanisms. Since all agents are assumed to be risk neutral, there is no loss of generality in considering only deterministic mechanisms (see Maskin, 1981).

[11] One can think of situations where the buyer is better informed about the probability of entry. Then we have a classic self-selection problem and all the results obtained in this section would also apply to this case.

offering a contract with lower liquidated damages, P_0. Such a contract would reduce the buyer's switching cost and could only profitably be offered by a seller facing a low probability of entry. We show however, that under certain conditions, signaling through the length of the contract is strictly better than signaling through liquidated damages.

To keep the analysis simple, we shall assume that the probability of entry is either "high" or "low." The incumbent knows the true probability but the buyer does not. Furthermore, as in Section I, the incumbent makes the contract offer. The situation described here is akin to an "informed Principal" problem (see Roger Myerson, 1983, and Maskin and Jean Tirole, 1985).

As in Section I, we shall assume that the entrant's costs are uniformly distributed on $[0,1]$. The incumbent's cost, on the other hand, is either $c=\frac{1}{2}$ or $c=k$, where $0<k<\frac{1}{2}$. Then the probability of entry is low when $c=k$ and it is high when $c=\frac{1}{2}$, since when $c=k$, we have

$$\underline{\phi} \equiv Pr(c_e \leq k) = k < \tfrac{1}{2}, \tag{8}$$

and when $c=\frac{1}{2}$, we have

$$\bar{\phi} \equiv Pr(c_e \leq \tfrac{1}{2}) = \tfrac{1}{2}. \tag{9}$$

The buyer's prior beliefs about the incumbent's costs are given by $m = Pr(c=k)$.

Under asymmetric information, it is no longer true that the seller can restrict himself with no loss to simple contracts, $c=\{P, P_0\}$. In fact, we show in our earlier paper that the incumbent seller can achieve the symmetric information optimal outcome by offering contracts of the form $c=\{P, P^e, P_0\}$ where P_0 is defined as in the previous section, P is the price the buyer pays if he trades with the incumbent and entry did not occur and P^e is the price the buyer pays if he trades with the incumbent and entry took place. Alternatively, when the incumbent only offers contracts of the form $c=\{P, P_0\}$, he can never attain the symmetric information optimal outcome. Thus simple contracts $c=\{P, P_0\}$ are suboptimal under asymmetric information. Thus, if the more general contracts $c=\{P, P^e, P_0\}$ are feasible asymmetric information puts no restrictions on the nominal length of the contract.

We give the following argument for why such contracts may not be feasible: First, "entry" may be a very complicated event to describe, when a firm can enter with a nonhomogeneous good. The incumbent must then decide what commodities qualify as "entrants" and, even if a list of such commodities can be defined, an entrant would have an incentive to produce a good which is not on that list whenever $P>P^e$. Alternatively, if $P^e>P$, there would be an incentive for the incumbent to claim that entry has occurred whenever there is an ambiguity about the event of entry. In short, the event of entry may be difficult to observe, let alone to verify.

Second, when $P>P^e$, the buyer could bribe someone to "enter" only to force the incumbent to lower his price. Vice versa, when $P<P^e$, the incumbent may want to bribe someone to enter.

When only simple contracts $c=\{P, P_0\}$ are feasible, asymmetric information can put restrictions on both the liquidated damages P_0, and the length of the contract. In the present model, contract length is somewhat artificially defined since production and trade take place only once. It should however be clear from what follows that the conclusions reached here carry over to a model with N periods of production and trade $(N \geq 2)$ where entry can take place in any of these N periods.

Here we compare the asymmetric information-contracting solution with the no-contracting solution and show that when the difference between high and low costs is sufficiently large, the low-cost incumbent is better off not signing a contract and leaving options open until the entry decision is taken by the potential competitor. In a model with N periods, this result would be modified and the low-cost incumbent would be better off signing a *shorter* contract than the high-cost incumbent.

When the seller makes a contract offer $c=\{P, P_0\}$, he conveys information about his type, so that the buyer's beliefs change.

Let the buyer's posterior beliefs be

$$\beta(c) = Pr(\phi = \bar{\phi}/c). \tag{10}$$

The buyer will only accept the contract if

$$1 - P \geq \beta(c)\bar{\phi}/2 + (1 - \beta(c))\underline{\phi}(1-k) \tag{11}$$

From (8) and (9) we can rewrite (11) as

$$1 - P \geq (\beta(c)/4) + (1 - \beta(c))k(1-k) \tag{12}$$

When the incumbent signs a contract $c = \{P, P_0\}$, the probability of entry is given by

$$Pr(c_e \leq P - P_0) = P - P_0. \tag{13}$$

Thus, the incumbent's payoff when he is respectively of type $\bar{\phi}$ or $\underline{\phi}$ is given by

(14)

$$V(c,\bar{\phi}) = (P - P_0)(P_0 - P + \tfrac{1}{2}) + P - \tfrac{1}{2}$$

$$V(c,\underline{\phi}) = (P - P_0)(P_0 - P + k) + P - k$$

for $P > P_0$, (otherwise $V(c,\bar{\phi}) = P - \frac{1}{2}$ and $V(c,\underline{\phi}) = P - k$). It is straightforward to verify that the Spence-Mirrlees condition is satisfied:

$$d/dk[-\partial V/\partial P/\partial V/\partial P_0] < 0. \tag{15}$$

In other words, it is more costly for an incumbent facing a higher probability of entry to lower P_0 than it is for an incumbent facing a lower probability of entry. Given condition (12) we can draw Figure 1 where $\bar{c}^* = \{P = \frac{3}{4}; P_0 = \frac{1}{2}\}$ is the optimal symmetric information contract when $\phi = \bar{\phi}$. Notice that this contract will always be accepted by the buyer since the right-hand side in (12) is increasing in β and when $\beta = 1$ (12) becomes

$$1 - P \geq \tfrac{1}{4}. \tag{16}$$

In addition, the contract $\bar{c}^*$ is the best con-

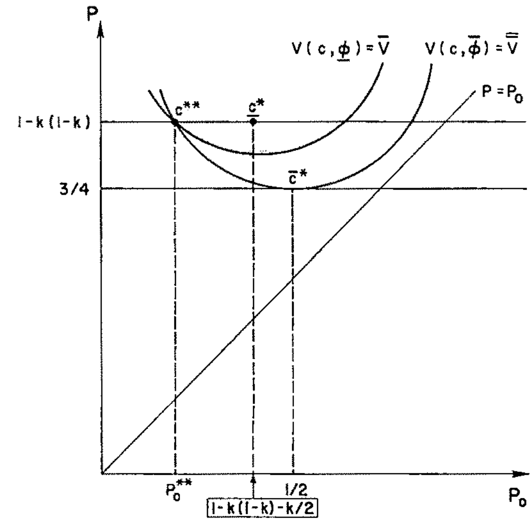


FIGURE 1

tract for the high-cost incumbent, among the class of contracts which generate beliefs $\beta(c) = 1$. It is common in signaling models to obtain a plethora of equilibria and our model is no exception to this rule. Any pair of contracts $(c, \bar{c}^*)$ where c is such that $P = 1 - k(1-k)$ and $0 \leq P_0 \leq P_0^*$ (see Figure 1) constitutes a separating equilibrium. Furthermore, any point in the shaded area in the diagram may be a pooling or semiseparating equilibrium of the signaling game. Following David Kreps (1984), however, we can refine the Bayesian equilibrium concept by using dominance and stability arguments and thus single out the best separating equilibrium $(c^{**}, \bar{c}^*)$ where c^{**} is defined as $c^{**} = \{P = 1 - k(1-k); P_0 = P_0^*\}$. How is P_0^* determined? It is the solution to the equation

$$V(c^{**}, \bar{\phi}) = V(\bar{c}^*, \bar{\phi}),$$

which can be rewritten as

$$(P - P_0)(P_0 - P + \tfrac{1}{2}) + P - \tfrac{1}{2} = \tfrac{1}{16} + \tfrac{1}{4}, \tag{17}$$

where $P = 1 - k(1-k)$.

Now P_0^* is the smaller root of this quadratic equation (see Figure 1) and is given

by

$$(18)\quad P_0^* = \left(\left(2P-\tfrac{1}{2}\right)-\sqrt{4P-3}\right)/2.$$

How does the optimal contract for the low-cost incumbent under asymmetric information compare with the optimal symmetric information contract given by $\underline{c}^* = \{P = 1-k(1-k);\ P_0 = (2P-k)/2\}$?

The optimal contract under asymmetric information, c^{**} specifies the same price P as c^*, but it specifies lower liquidated damages: $P_0^* < P_0$. It is straightforward to compute that $P_0^* < P_0$ reduces to

$$(19)\qquad 1+4k^2 > 5k-\tfrac{1}{2}.$$

And for all $0<k<\frac{1}{2}$ this inequality is verified.

Intuitively, the incumbent with low costs signals his type by offering to reduce liquidated damages below the first-best level. His information is credibly transmitted since it is too costly for the high-cost incumbent to reduce P_0 to that level and thereby induce too much entry.

We now show that for small k, the low-cost incumbent is better off not signing a contract than signing c^{**}. If the low-cost seller does not sign a contract, his expected profits are given by

$$(20)\qquad (1-\underline{\phi})(1-k)=(1-k)^2.$$

If he signs c^{**} he gets

$$(21)\quad V(c^{**},\underline{\phi}) = (P-P_0^*)(k-(P-P_0^*))+P-k,$$

where $\quad P=1-k(1-k)$

and $\quad P-P_0^* = \frac{1}{4}+\frac{1}{2}\sqrt{4(1-k(1-k))-3}\,.$

It remains to show that for small k, we have

$$(22)\quad \left[\tfrac{1}{4}+\tfrac{1}{2}\sqrt{4(1-k(1-k))-3}\,\right]k - \left[\tfrac{1}{4}+\tfrac{1}{2}\sqrt{4(1-k(1-k))-3}\,\right]^2 + 1-k(1-k)-k \le (1-k)^2$$

And (22) reduces to

$$(23)\quad k \le \tfrac{1}{4}+\tfrac{1}{2}\sqrt{4(1-k(1-k))-3}$$

which is clearly verified for small k. Also, for k close to $\frac{1}{2}$, (23) is not satisfied. We summarize the above discussion in the following proposition:

PROPOSITION 1: *Under asymmetric information about the probability of entry (or equivalently about the incumbent's costs), the optimal contracting solution is such that*

(a) *the high-cost incumbent signs the optimal symmetric information contract* $\bar{c}^* = \{P = \frac{3}{4};\ P_0 = \frac{1}{2}\}$.

(b) *the low-cost incumbent either signs the second-best contract*

$$c^{**} = \left\{ P = 1-k(1-k);\quad P_0^* = P-\tfrac{1}{4}-\tfrac{1}{2}\sqrt{4P-3}\,\right\}$$

(when k is close to $\frac{1}{2}$) or does not sign a long-term contract at all (when k is close to zero).

(c) *c^{**} is characterized by the property that liquidated damages (P_0^*) are lower than in the optimal symmetric information contract,*

$$\underline{c}^* = \{P = 1-k(1-k);\ P_0 = P-(k/2)\}.$$

One can explain Proposition 1(b) as follows. As k becomes smaller the price $P = 1-k(1-k)$ rises, which makes it more attractive for the high-cost firm to mimic the low-cost firm's behavior. In order to discourage the high-cost firm from cheating, the low-cost firm must therefore increase the gap $P-P_0 = [\frac{1}{4}+\frac{1}{2}(4(1-k(1-k))-3)^{1/2}]$. But this is equivalent to raising the probability of entry after a contract has been signed (see equation (13)). There comes a point where $\phi' = P-P_0 \ge \underline{\phi} = k$; that is, by raising $P-P_0$, the low-cost firm raises the *ex post probability of entry* ($\underline{\phi}'$) above the *ex ante probability of entry* ($\underline{\phi}$) (see (23)). This essentially involves subsidizing some inefficient entrants to enter the market. The incumbent

then gets a negative transfer from the entrant. He can do strictly better by not offering any transfer (i.e., by not signing a contract at all).

We have thus established that the *nominal* length of the contract may serve as a signal of the probability of entry. This result confirms the following basic intuition:

The buyer reasons as follows when he is offered a contract: "If the incumbent wants to sign a contract of a long duration he must be worried about entry, so that I infer from this that the probability of entry is high and I will only accept to sign this contract if he charges a low price. If, on the other hand, the incumbent offers a short-term contract, he reveals that he is not much preoccupied about entry, so that I will be willing to accept a higher price."

The result obtained in Proposition 1(c) implies that the social cost is smaller in the asymmetric information case than in the symmetric information case. That is, liquidated damages (P_0^*) are smaller in c^{**} than in $\underline{c}^*$; therefore fewer efficient firms will be kept out of the market. It is worth emphasizing this point, since one usually thinks of asymmetric information as a constraint that prevents agents from reaching a socially efficient outcome (a first-best optimum). This is a general theme in Agency theory (see Oliver Hart and Holmström, 1985). Here, on the contrary, asymmetric information about the incumbent's costs may actually force agents to choose the socially efficient outcome (whenever the condition in (23) is verified). *The informational asymmetry constrains the monopoly power of the incumbent and the buyer with respect to the entrant.* There is another interpretation of this result. Remember that the incumbent and the buyer are constrained in the first place by the informational asymmetry about the entrant's costs. Then, the conclusion reached here is that if there exists another informational asymmetry between the buyer and the incumbent (about the latter's cost) *the two informational constraints may cancel each other out.*

This is an important observation for agency theory. Informational constraints do not necessarily add up; they may cancel out.

III. Optimal Contracts with Several Buyers

One may wonder to what extent the results obtained in Sections I and II depend on the assumption that there is only one incumbent seller and one buyer? This section attempts to give a partial answer to this question. We compare in turn the situation where there is one buyer but several incumbent sellers, and the situation where there is one incumbent seller but several buyers. All the results established in Section I are valid in each case. Moreover, new interesting features are introduced in the latter situation, where a single incumbent negotiates with several buyers.

Consider first the situation where there are two or more identical sellers but only one buyer. Then, Bertrand competition essentially gives all the bargaining power to the buyer; he gets all of the surplus but the form of the optimal contract does not change. The buyer sets P_0 in the same way as the seller does, when the seller makes the contract offer.

The interesting situation is when there are several buyers and one seller. In this case, the entrant's profits depend on how many customers he can serve in the post-entry game. What is crucial, however, is how the size of the entrant's potential market affects the probability of entry. If the probability of entry is independent of the size of the market, then the case of several buyers reduces to the case of one buyer. In general, however, the size of the market will affect the probability of entry. For example, if the entrant must pay a fixed cost of entry, then his average cost is decreasing in the number of customers served and the probability of entry is increasing in the number of customers.

In this latter case, when one buyer signs a long-term contract with the incumbent, he imposes a negative externality on all other buyers. By locking himself into a long-run relation with the seller, he reduces the size of the entrant's potential market so that, *ceteris paribus*, the probability of entry will be smaller. As a result, the other buyers will have to accept higher prices. We show that the incumbent can exploit this negative ex-

ternality to extract more (possibly all) surplus out of each buyer. In some cases, the seller can impose the monopoly price ($P=1$) on each buyer, even though the *ex ante* probability of entry is arbitrarily close to one (*ex ante* refers to the no-contract situation). In addition, the seller can extract part of the entrant's surplus by choosing damages (P_0) appropriately, so that we get the paradoxical result that a seller facing a threat of entry may be better off than a natural monopoly. To reach this conclusion, we must push the logic of the game to its limits. This result is thus interesting mainly for illustrative purposes.

We will only consider the case of two buyers and one seller.[12] Both buyers are identical and have a reservation price $P=1$. The incumbent is as described in Section I. The entrant has the same unit costs as in Section I; in addition, he may face a fixed cost of entry, $F \geq 0$. We shall first consider the problem where F is strictly positive. Then, in the absence of any contract, the entrant's profit is given by

$$\pi_e = 2\left(\tfrac{1}{2} - c_e\right) - F. \tag{24}$$

Thus, the *ex ante* probability of entry is given by

$$\phi = Pr(\pi_e \geq 0) = (1-F)/2. \tag{25}$$

Suppose now that one of the buyers signs a contract with the incumbent where $P_0 = +\infty$. Then in the post-entry game, this buyer will never switch to the entrant. The latter can now hope to get at most:

$$\hat{\pi}_e = \tfrac{1}{2} - c_e - F. \tag{26}$$

The other buyer therefore faces a lower likelihood of entry given by

$$\hat{\phi} = Pr(\hat{\pi}_e \geq 0) = (1-2F)/2. \tag{27}$$

More generally, whenever one buyer signs a contract with the incumbent of the form $c=\{P, P_0\}$, the other buyer faces a new probability of entry given by

$$\hat{\phi} = \max\left\{\frac{P - P_0 + \frac{1}{2} - F}{2}; \frac{1-2F}{2}\right\}. \tag{28}$$

We will analyze the negotiation game where the incumbent makes simultaneous contract offers to both buyers. The case where the incumbent makes sequential offers is considered in our earlier paper. There we establish that the timing of offers does not matter. The same outcome is obtained in the simultaneous offers case as in the sequential offers case.

The incumbent can without loss restrict the set of contracts to be of the form $c=\{P, P_0, P^r, P_0^r\}$, where

P = the price a buyer must pay if he trades with the incumbent and the other buyer has signed a long-term contract;

P_0 = the damages a buyer must pay if he switches to the entrant and the other buyer has signed a contract with the incumbent;

P^r = the price a buyer must pay if he trades with the incumbent and the other buyer did not sign a contract;

P_0^r = the damages a buyer must pay if he trades with the entrant and the other buyer did not sign a contract with the incumbent.

It is implicitly assumed here that all contracts are publicly observable. This is a strong assumption. In practice, all contracts are not observable. As a result, one can never be certain when a contract is observed, whether there does not exist a hidden contract which cancels the effects of the observed contract. In our model, however, the incumbent has an incentive to publicize all of his contracts, as will become clear below. Thus, hidden contracts are not a problem.

When the seller makes a contract offer $c=\{P, P_0, P^r, P_0^r\}$ to each buyer, B_1 and B_2, the latter play a noncooperative game where they have two pure strategies: "accept" and "reject." The payoff matrix of this game is represented in Table 1. By choosing P^r and P_0^r appropriately, the incumbent can ensure that $\hat{\phi} = (1-2F)/2$.

[12] We deal with the generalization to n buyers ($n \geq 2$) in our earlier paper.

TABLE 1—B_1

	Accept	Reject
Accept	$1-P$, $1-P$	$1-p^r$, $\hat{\phi}/2$
Reject	$\hat{\phi}/2$, $1-p^r$	$\phi/2$, $\phi/2$

Essentially, this involves choosing P_0^r large enough so that the buyer who accepted a contract will not switch to the entrant. Now, accept is a (weakly) dominant strategy when

$$(29) \qquad 1-P \geq \hat{\phi}/2 = (1-2F)/4;$$

$$(30) \qquad 1-P^r > \phi/2.$$

When the incumbent offers a contract to both buyers such that (29) and (30) are satisfied (and such that $\hat{\phi} = (1-2F)/2$), the unique Nash equilibrium is for both buyers to accept the contract offer. As a result, both buyers receive a strictly lower payoff in equilibrium than if they both rejected the contract, since $\hat{\phi} < \phi$.

Thus when there are several buyers contracting with the incumbent, there is another reason why rational buyers are willing to perpetuate the monopoly position of the seller. As Steven Salop puts it, contracts "...are valued by each buyer individually even while they create an external cost to all other buyers" (1986, p. 273). He calls this situation a "*free-rider effect in reverse*" (emphasis added).

In addition to this effect, the seller can set P_0 appropriately so as to extract the maximum expected surplus from the entrant. To summarize, in this simple model with simultaneous offers, the set of optimal contracts is given by

$$(31) \qquad c^* = \left\{ P = 1 - \frac{(1-2F)}{4}; \quad P_0 = P - \frac{F+1}{4}; \quad P^r < 1 - \frac{\phi}{2}; P_0^r > P^r + \tfrac{1}{2} + F \right\}.$$

And at the optimum the incumbent's expected payoff is given by

$$(32) \qquad \pi = \left(2(P-P_0) - F\right)\left(2\left(P_0 - P + \tfrac{1}{2}\right)\right) + 2\left(P - \tfrac{1}{2}\right) = \frac{(1-F)^2}{2} + 1 - \frac{(1-2F)}{2}.$$

Suppose now that $F \geq \frac{1}{2}$, then $\hat{\phi} = 0$ and the incumbent is able to impose the monopoly price $(P = 1)$ on the buyer. His expected payoff at the optimum is then given by

$$(33) \qquad \pi = \left((1-F)^2/2\right) + 1.$$

Thus the incumbent does strictly better than a natural monopoly, since he can also extract some of the potential entrant's surplus. On the other hand, when $F = 0$, we have $\hat{\phi} = \phi = \frac{1}{2}$, and the "free-rider effect in reverse" disappears, so that the two buyers case reduces to a one-buyer case, where the customer purchases two units rather than one. In other words, when the probability of entry is independent of the size of the market, competition among buyers does not matter.

Thus the principles established in the one buyer-one seller case remain valid when we allow for either more than one buyer or more than one seller. The analysis is somewhat incomplete since we did not deal with the several buyers-several sellers case. The results obtained in Section I carry through to this more general model (see Diamond-Maskin). As far as the results in this section are concerned, it is likely that sellers will not be able to exploit to the same extent the free-rider effect in reverse.

IV. Conclusion

The principles formalized in this paper are very general. What is basically required for contracts to constitute a barrier to entry is that post-entry profits for the incumbent in the absence of any contract be lower than pre-entry profits (and vice versa for consumers). In addition, it is necessary that the incumbent cannot discriminate between entrants of various levels of efficiency. This is a

rather mild assumption if one interprets the entrant's cost as an opportunity cost of entry as in our earlier paper. Throughout the paper we interpreted P_0 to be "liquidated damages," but P_0 may also represent down payments, deposits, collateral, future discounts, and benefits, etc. Thus, the analysis developed here has potentially a wide range of applicability.

Casual empiricism suggests that "endogenous switching costs" for customers are a widespread phenomenon. In the housing market, for example, advance deposits in rental contracts can be interpreted as serving this function (there are, of course, also moral hazard reasons for requiring deposits). Paul Klemperer (1986) provides a number of examples of endogenous switching costs, like frequent flyer programs, trading stamps, deferred rebates by shipping firms, etc. Also, fixed fees in franchise contracts may be used to extract some rent from a potential competitor. The contract between Automatic Radio Manufacturing Co. and Hazeltine Research (see *Automatic Radio Manufacturing Co. v. Hazeltine Research Inc.*, 1950) is a good example. Automatic Radio had to pay a fixed fee irrespective of whether it exploited the patents licensed by Hazeltine. Any new licensor therefore faced an entry barrier equal to the amount of this fee. Another striking example is the case of Bell Laboratories when it invented the transistor. There were other research institutes competing with Bell Laboratories. In order to preempt them, Bell Labs offered to publicize the technology to any potential licensee, in exchange for a fixed fee of $25,000. This fee served the same function as P_0, in the contract above. Moreover Bell Lab's strategy was to become the industry standard. Thus any individual licensee would have to take into account the additional switching cost of not being standardized (see E. Braun and S. Macdonald, 1978). Our analysis provides a rationale for the practices described here and explains why rational customers cooperate with firms in these anticompetitive practices. Unfortunately, the variety and potential complexity of these contractual clauses makes the task for antitrust authorities very difficult.

A rapidly growing literature on exogenous switching costs is related to our present study (see Klemperer for a recent thorough exposition). The welfare conclusions obtained in this research are radically different from ours. For example, in Klemperer, entry may be socially inefficient because consumers dissipate the gains from entry (in terms of lower prices and higher output) by incurring the socially wasteful switching costs. In our model, the social cost comes from insufficient entry; when entry occurs it is always welfare improving. Salop also studies the effect of various clauses, such as the "meeting the competition clause" or the "clause of the most favored nation" on competition. His emphasis is more on cartel coordination than entry prevention. In our model a "meeting the competition clause" would preclude entry since the entrant could never undercut the incumbent. *We have shown, however, that it is optimal not to eliminate entry completely*. Therefore, such clauses will never be adopted for entry-deterrence purposes; they may however be useful to facilitate cartel coordination, as Salop shows, since they increase the cost of price cutting.

Our theory of contract length is a substantial departure from existing theories. Most explanations have emphasized the idea that contract length is determined as a tradeoff between recontracting costs and the costs associated with the incompleteness of the contract (see Williamson, 1975, 1985; Ronald Dye, 1985a; Jo Anna Gray, 1976). A notable exception is Harris and Holmström. In practice, uncertainty about the future and the cost of writing complete contracts are without doubt important elements in the determination of contract length. The difficulty from a theoretical perspective is however that uncertainty about the future and "transaction costs" are notoriously vague categories. If contracts are to be incomplete what contingencies should the parties leave out of the contract? This is a very difficult question which has only received partial answers (see Dye, 1985b, and Hart-Holmström). Explanations of contract length based on contractual incompleteness crucially depend on how one answers this question (see Dye, 1985b). In this paper we have

sidestepped the difficulty to provide a story based on asymmetric information. We believe that signaling aspects are important in the determination of contract length and view our explanation as complementary to the existing theories.

Recently, Benjamin Hermalin (1986) has developed another theory of contract length based on asymmetric information. He considers a competitive labor market where initially workers have private information about productivity but where in a later stage this information becomes public (for example, through output observations). He shows that by varying contract length, it is impossible for firms to *profitably* screen out low-productivity workers from high-productivity workers. Ideally, a firm wants to retain only high-productivity workers, but long-term contracts are most attractive to low-productivity workers. Thus, by screening out workers, the firm achieves the opposite of what it wants: it offers long contracts to low-productivity workers and short contracts to high-productivity workers. In equilibrium, either firms offer only short-term contracts, or they offer "trivial" long-term contracts that replicate the outcome achieved with short-term contracts. In our model, on the contrary, signaling (or screening) works. Moreover, when it is optimal for the low-cost incumbent to sign a short-term contract, there does not exist an alternative trivial long-term contract. One can view our explanation and Hermalin's as dual: in his model the high-productivity sellers do not want to be locked in a long-term contract; here it is the buyer who does not want to forego future opportunities.

REFERENCES

Aghion, Philippe and Bolton, Patrick, "Entry-Prevention through Contracts with Customers," unpublished, 1985.

Blair, Roger D. and Kaserman, David L., *Law and Economics of Vertical Integration and Control*, New York: Academic Press, 1983.

Bork, Robert H., *The Antitrust Paradox*, New York: Basic Books, 1978.

Braun, E. and Macdonald, S., *Revolution in Miniature: The History and Impact of Semiconductor Electronics*, New York: Cambridge University Press, 1978.

Caves, Richard, E., "Vertical Restraints in Manufacturer-Distributor Relations: Incidence and Economic Effects," mimeo., Harvard University, 1984.

Dasgupta, Partha and Stiglitz, Joseph, "Sunk Costs and Competition," mimeo., Princeton University, 1984.

______, **Hammond, Peter and Maskin, Eric,** "The Implementation of Social Choice Rules: Some General Results on Incentive Compatability," *Review of Economic Studies*, April 1979, *46*, 185–206.

Diamond, Peter A. and Maskin, Eric, "An Equilibrium Analysis of Search and Breach of Contract, I: Steady States," *Bell Journal of Economics*, Spring 1979, *10*, 282–316.

Dixit, Avinash, "A Model of Duopoly Suggesting a Theory of Entry-Barriers," *Bell Journal of Economics*, Spring 1979, *10*, 20–32.

______, "The Role of Investment in Entry Deterrence," *Economic Journal*, March 1980, *90*, 95–106.

Dye, Ronald, (1985a)"Costly Contract Contingencies," *International Economic Review*, February 1985, *26*, 233–50.

______, (1985b)"Optimal Length of Labor Contracts," *International Economic Review*, February 1985, *26*, 251–70.

Fudenberg, Drew and Tirole, Jean, "Dynamic Models of Oligopoly," in J. Lesourne and H. Sonnenschein, eds., *Fundamentals of Pure and Applied Economics*, New York: Harwood Academic Press, 1986.

Gray, Jo Anna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, *2*, 221–35.

Harris, Milton and Holmström, Bengt, "On the Duration of Agreements," mimeo., IMSSS, Stanford University, 1983.

Hart, Oliver and Holmström, Bengt, "The Theory of Contracts," in T. Bewley, ed., *Advances in Economic Theory*, New York: Cambridge University Press, 1985.

______ **and Moore, John,** "Incomplete Contracts and Renegotiation," mimeo., MIT 1985.

Hermalin, Benjamin, "Adverse Selection and Contract Length," mimeo., MIT, 1986.

Klemperer, Paul, "Markets with Consumer Switching Costs," unpublished doctoral dissertation, Graduate School of Business, Stanford University, 1986.

Kreps, David M., "Signaling Games and Stable Equilibrium," mimeo., Stanford University, 1984.

Maskin, Eric, "Randomization in Incentive Problems," mimeo., 1981.

______ **and Tirole, Jean,** "Principals with Private Information, II: Dependent Values," lecture notes, 1985.

Myerson, Roger B., "Mechanism Design by an Informed Principal," *Econometrica*, November, 1983, *51*, 1767–97.

Posner, Richard A., *Antitrust Law: An Economic Perspective*, Chicago: University of Chicago Press, 1976.

Salop, Steven, "Practices that (credibly) Facilitate Oligopoly Coordination," in J. Stiglitz and F. Mathewson, eds., *New Developments in the Analysis of Market Structure*, Cambridge: MIT Press, 1986.

Spence, A. Michael, "Entry, Capacity, Investment and Oligopolistic Pricing," *Bell Journal of Economics*, Autumn 1977, *8*, 534–44.

Stiglitz, Joseph and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–409.

Williamson, Oliver E., "The Vertical Integration of Production: Market Failure Considerations," *American Economic Review*, March 1971, *61*, 112–23.

______, *Markets and Hierarchies: Analysis and Antitrust Implications*, New York: Free Press, 1975.

______, "Assessing Vertical Market Restrictions: Antitrust Ramifications of the Transaction-Cost Approach," *University of Pennsylvania Law Review*, April 1979, *127*, 953–93.

______, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

Automatic Radio Manufacturing Co. v. Hazeltine Research Inc., 339 U.S. 827, 834, 1950.

United States v. United Shoe Machinery Corporation, 258 U.S. 451, 1922.

$R\&D$ Rivalry with Licensing or Imitation

By MICHAEL L. KATZ AND CARL SHAPIRO*

We study the rivalry between two firms to develop an innovation in a dynamic setting that allows for postdevelopment dissemination of the innovation, such as licensing or imitation. This dissemination may cause the noninnovating firm to benefit from the discovery. When this occurs, conventional results in the economics of R&D no longer need apply. We find that industry leaders will tend to develop minor innovations, but will develop major innovations only if imitation is difficult.

Technological progress is the driving force behind long-run economic performance. The pace of innovation in market economies depends, in turn, upon private firms' incentives to innovate. When there is only a single firm capable of undertaking a given research and development ($R\&D$) project, that firm compares the profits it would earn were it to undertake the project with the profits it would obtain if no one were to innovate. We call this difference in the firm's profits its *stand-alone incentive* to develop the innovation.

A more common and important situation is one in which there is rivalry to innovate. Consider the case of two firms engaged in $R\&D$ competition. Each firm's incentives to innovate depend crucially on the actions of its rival. If neither firm has innovated, then each firm uses its existing technology and earns some baseline flow of profits. If firm i believes that its rival will not innovate, then firm i compares its post-innovation profits with its baseline profits; the firm is driven by its stand-alone incentive. But a firm may believe that its rival will innovate if it does not. In this case, firm i's incentives to innovate are driven by the difference between its profits as the "winner" of the $R\&D$ competition and its profits as the "loser," where its rival obtains the innovation first. We call this difference firm i's *preemption incentive*.

If each firm suffers a reduction in profits when its rival innovates, the preemption incentives are larger than the stand-alone incentives. In this case, the winner must be the firm with the greater preemption incentive, and $R\&D$ competition takes the form of a *race* to be the innovator. There is now a large literature in which the authors examine private incentives to innovate in this setting—excellent examples include Partha Dasgupta and Joseph Stiglitz (1980), Richard Gilbert and David Newbery (1982), and Jennifer Reinganum (1982). The main finding of this literature is that the firms who are most successful using the baseline technology will tend to be the developers of the new technology as well.

There are cases, not considered by earlier authors, in which a firm benefits from its rival's development of the innovation; the profits from "losing" exceed the baseline profits. In these cases, the stand-alone incentives exceed the preemption incentives, and the nature of $R\&D$ competition may become that of a *waiting game* rather than a race.[1]

*Department of Economics and Woodrow Wilson School, Princeton University, Princeton, NJ, 08544. This paper is an abridged version of our 1984 paper, "Perfect Equilibrium in a Development Game with Licensing or Imitation." The interested reader is referred to that paper for a more complete technical analysis. This material is based on work supported by the National Science Foundation under grants no. SES-8209412 (Katz) and SES-8408622 (Shapiro). We both thank the Alfred E. Sloan Foundation for financial support as Sloan Research Fellows. Shapiro also thanks the Institute for International Economic Studies, University of Stockholm, where the paper was revised. Comments by Avinash Dixit, Jean Tirole, and two referees are appreciated.

[1] Our waiting games are quite different from those studied in the existing literature. That literature has emphasized symmetric mixed strategies in the case of

Moreover, the loser may wish to lose as soon as possible, since the developing firm is providing a public good for the industry. When the noninnovating firm benefits from a discovery, the stand-alone incentives may determine the development date and the identity of the innovating firm. As we will show, the firm with the lower baseline profits under the old technology may be the innovator.

Why would a firm ever prefer innovation by its rival to the status quo?[2] In the case of perfect patent protection and drastic innovations studied by earlier authors, it would not. But when patents are less than perfect, imitation may occur. And when the innovation is not essential to survival in the industry, licensing between direct competitors may occur even if patent protection is perfect. These forms of post-innovation dissemination give rise to the possibility that a firm benefits from innovation by its rival.

Consider first the significance of imitation. If a noninnovating firm can imitate quickly, effectively, and at low cost, it may benefit from a discovery even when the innovator is a product-market competitor. Imitation is a common occurrence. Edwin Mansfield, Mark Schwartz, and Samuel Wagner (1981) found that about 60 percent of the patented successful innovations in their sample were imitated within four years.[3] In their survey of industry *R&D* personnel, Richard Levin et al. (1986) found that even major patented innovations could be imitated within three or fewers years in well over half of the 129 lines of business covered. Levin et al. also found that *R&D* managers generally placed relatively little faith in the ability of patents to prevent rivals from imitating their product or process innovations. In fact, managers placed greater faith in secrecy as a means of protecting their property rights.

The essential features of imitation are two. First, the imitator makes no payment to the innovator. Second, the development costs incurred by the imitator typically are lower than those of the innovator. Mansfield et al. (1981) found that the imitator's costs were on average only 65 percent of the innovator's costs.[4] Levin et al. found that in over 80 percent of the lines of business covered in their survey, the imitation costs for an unpatented "typical" innovation were less than 75 percent of the original innovation costs. This relationship held for patented typical innovations in roughly 65 percent of the lines of business.

Even when patents are sufficiently strong to block imitation, dissemination of an innovation may take place through licensing, where owners of intangible assets receive payments from other firms in return for divulging their secret technical know-how or sharing patents. Licensing of both patented and unpatented property is common in technologically progressive industries. Some sense of licensing's importance can be gleaned from the fact that in 1984 alone the licensing and sale of intangible property by American corporations to foreign firms (including subsidiaries of American firms acting as licensees) accounted for almost $8 billion of revenues.[5]

complete information (the "war of attrition") and pure strategies in the case of incomplete information. See Barry Nalebuff and John Riley (1984) and Drew Fudenberg and Jean Tirole (1986), for example. We look at pure strategies in an asymmetric, nonstationary, complete information setting.

[2]A firm is likely to benefit from another's innovation if the two produce complementary products, or if the innovator is a supplier or customer of the first firm. Our theory is general enough to encompass such cases, but we emphasize *R&D* rivalry between firms that are direct competitors.

[3]Mansfield et al. (1981, p. 909), surveyed 48 product innovations in the U.S. chemical, drug, and electronics and machinery industries, of which roughly 70 percent were patented.

[4]The estimate of the relative size of imitation costs compares the patentee's development costs with an imitator's copying costs. To the extent that the patentee is more efficient at development, this underestimates the costs that the imitator would have incurred to develop on its own.

[5]*Survey of Current Business*, March 1985, p. 41. This number understates the true value of trade in intangibles to the extent that the U.S. corporate income tax motivates companies to set artificially low transfer prices in selling to their foreign subsidiaries. Although current data are not available for the dollar volume of domestic licensing, Robert Wilson (1977) reports that domestic licensing receipts accounted for 44 percent of total

Although important, the effect of licensing and imitation on development incentives have been largely ignored in the literature on dynamic $R\&D$ rivalry.[6] In this paper, we study the rivalry between two firms to develop an innovation in a dynamic model that has a payoff structure general enough to encompass the existence of licensing and imitation. Our framework allows us to characterize the pace and nature of $R\&D$ rivalry more generally than has been done by earlier authors. Our theory suggests that the issue of whether a currently dominant firm or a currently small firm will develop an innovation depends upon three characteristics of the innovation: 1) its overall size; 2) the sensitivity of the cost reduction engendered by the innovation to the innovator's initial cost level; and 3) the extent to which the innovation can be imitated.

Consider an innovation that incrementally adds to the existing technology and, because it is used in conjunction with current technology, permits approximately equal cost reductions for all firms. We call such an innovation *minor*. In contrast, a *major innovation* is one that permits a large cost reduction by replacing the existing technology, and which tends to equalize post-innovation costs (i.e., preexisting technological differences across firms are largely irrelevant when using the new technology).

Our theory generates two testable hypotheses. First, the firm with the higher baseline profits, the industry leader, will tend to develop major innovations if and only if imitation is difficult. In contrast, the leader will tend to develop minor innovations whether or not imitation and licensing are feasible. The limited data that we have found are consistent with this pattern of development.

Our framework also allows us to explore the effects of changes in the ease of licensing and imitation on the timing of innovation. This analysis sheds light on important issues of antitrust and patent policy. For example, does antitrust policy that makes it easier for firms to license (say, by letting them communicate more fully) hasten technological progress? Such a reduction in the technology transfer costs raises the licensor's profits and might be expected to create incentives for more rapid innovation. But the cost reduction actually may delay development (for empirically supported values of the critical parameter) because each firm is more content to wait, lose the race, and become a licensee. We also find that, while imitation typically delays development, it does so for an unexpected reason.

The paper is organized as follows. Section I describes a general framework of $R\&D$ competition, and gives a reduced-form characterization of the equilibria. In Section II, this general framework is applied to a series of detailed examples. These examples are used to relate the equilibrium date of innovation and the identity of the developer to such underlying elements of market structure as the size of the innovation, the pre-innovation configuration of costs, and the licensing and imitation possibilities. We derive comparative statics results in Section III that allow us to study the effects of antitrust and patent policy. There is a short concluding section.

I. $R\&D$ Rivalry: A General Framework

A. *The Model*

Technological progress is brought about by a variety of research and development activities. While such activities are often lumped together as "$R\&D$," there are in fact important differences between "R" and "D."

licensing receipts by U.S. firms in 1971, and Michael Rostoker (1984) found that firms in his sample tended to license twice as often to domestic firms as to foreign firms during 1975–80. Rostoker also found that 68 percent of the licensing agreements involved at least some information that was not subject to patent protection (i.e., know-how and trade secrets).

[6] Reinganum (1981b) allows for spillovers *during* the $R\&D$ race, but not licensing or imitation of the innovation itself. While Reinganum (1982) analyzes imitation, she maintains the crucial assumption that a firm always prefers winning to losing (so a waiting game cannot arise). Nancy Gallini (1984), Gallini and Ralph Winter (1985), Morton Kamien and Yair Tauman (1986), Katz (1986), Katz and Shapiro (1985, 1986), and A. Michael Spence (1984) all study various aspects of licensing and imitation, but do so in timeless settings, not in the context of dynamic $R\&D$ rivalry. Finally, F. M. Scherer (1980) provides an informal discussion of the fact that imitation could give rise to a waiting game.

Basic research discoveries often are difficult to appropriate, and frequently are undertaken outside of the for-profit sector. Development efforts, on the other hand, often generate benefits that can be largely appropriated via intellectual property rights, and are undertaken chiefly by for-profit firms.

We consider competition between two firms to put an innovation into practice. This development competition takes place in an explicitly dynamic environment. We assume that basic research activities are undertaken nonstrategically by government and university scientists and engineers, and that basic research results are unappropriable. Basic research gradually reduces the costs that the firms must bear to develop a given innovation. Unlike basic research, development is rapid and at least partially appropriable. We permit a "strong" system of property rights for development results, which tends to make patent licensing feasible and imitation difficult. We also permit a "weak" system, which may give rise to the imitation of development results. In fact, one of the key questions that we address is how varying degrees of "strength" affect development incentives.

At each point, the firms make production decisions using their current technologies. The flow of profits earned by each firm depends in general upon demand conditions, duopolistic behavior, and the production costs at the two firms. In Section II below, we shall express the firms' profits in terms of these underlying variables. In order to focus on the *R&D* rivalry, we initially leave the flow profits in reduced form. We denote by π_i^0, $i=1,2$, firm i's flow of profits prior to the development of the innovation, during which time each firm produces using the old, or in-place, technology. Our framework is general enough to allow for one firm's being a monopolist prior to the innovation's development. If the innovation never is developed, then firm i earns π_i^0 forever, for a total payoff of π_i^0/r, where $r>0$ is the interest rate.[7]

In addition to its production activities, each firm must choose whether to expend the resources to develop a new product or process. Development decisions are made only at discrete dates spaced Δ apart. Thus, at each date $t=0,\Delta,2\Delta,\ldots$, each firm decides whether to wait for basic researchers to make further progress that will lower the costs of development, a strategy that we call "wait," or to develop the innovation using existing knowledge, which we call "develop." We assume that Δ is close to zero; indeed, we are interested in the results as $\Delta\to 0$.[8]

As soon as one firm selects the strategy develop, the firms have no further opportunities to improve their technologies.[9] Once chosen, development is assumed to take place instantaneously. If firm i is the first to choose develop and does so at time T, then we say that "firm i wins at T." The current cost incurred by the developing firm if it innovates at date T is denoted by $K(T)e^{rT}$, where $K(T)$ is the present value, viewed at time $t=0$, of development expenses.

We assume that development initially is too expensive to be attractive.[10] Over time, however, basic research improves the underlying technology and reduces the current costs of development (i.e., $d(K(T)e^{rT})/dT<0$). We assume that development costs fall at a decreasing rate (i.e., $d^2(K(T)e^{rT})/dT^2>0$).[11] Denote by $K\geq 0$ the limit (as $T\to\infty$) of current development costs.

When firm i innovates, it introduces the new technology into its production activities.

[7]Throughout the paper, subscripts denote the firm in question and superscripts refer to the identity of the innovating firm. A superscript of naught denotes a pre-innovation variable.

[8]For a more complete formal analysis of the limiting arguments as $\Delta\to 0$, see our 1984 working paper.

[9]This structure is one way in which our analysis differs from that of technology adoption (as in Reinganum, 1981a, or Fudenberg and Tirole, 1985), where a second firm may later adopt the new technology on the same terms as the first firm.

[10]In terms of the notation below, we require $K(0)\geq\pi_i^i/r$, $i=1,2$.

[11]This last assumption ensures that the firms' payoffs are quasi concave in the development date. Our development cost assumptions are much like those made by Fudenberg and Tirole (1985) in the context of the adoption of new technologies. Our analysis has some similarities with that paper. The key difference is that their adoption game cannot be a waiting game, since each firm is hurt when its rival adopts the superior technology.

At the same time, its rival's technology may change, due either to imitation or licensing. We denote by π_i^i firm i's flow profits subsequent to its own development of the innovation. After firm i has innovated, firm j's flow profits become π_j^i. We consider both cases in which $\pi_j^i < \pi_j^0$ and cases in which $\pi_j^i > \pi_j^0$. Industry profits when firm i has developed are given by $\pi^i \equiv \pi_i^i + \pi_j^i$.[12]

If firm i wins at T, it enjoys flow profits of π_i^0 for $t < T$, flow profits of π_i^i for $t \geq T$, and incurs the cost $K(T)$. Hence, firm i's payoff if it wins at T is

$$W_i(T) = \int_0^T \pi_i^0 e^{-rT} dt + \int_T^\infty \pi_i^i e^{-rt} dt - K(T),$$

or

$$(1) \quad W_i(T) = \frac{1-e^{-rT}}{r}\pi_i^0 + \frac{e^{-rT}}{r}\pi_i^i - K(T), \quad i=1,2.$$

Similarly, the payoff to firm i if firm j wins at T (i.e., if i loses at T), is

$$(2) \quad L_i(T) = \frac{1-e^{-rT}}{r}\pi_i^0 + \frac{e^{-rT}}{r}\pi_i^j, \quad i=1,2 \; j \neq i.$$

The only remaining outcome is that both firms choose develop for the first time simultaneously. In this case, we assume that a coin is flipped determining the identity of the winner, so that firm i earns $\{W_i(T)+L_i(T)\}/2$.[13]

[12] We need not assume that the firms' flow profits are constant after the innovation. Indeed, with imitation lags, they will not be. All that we require is that the present value of firm j's flow of profits, measured as of the date that firm i innovates, be π_j^i/r.

[13] This formulation implicitly assumes that the firm losing the coin flip need not incur the development costs. In our 1984 paper, we prove that, so long as an equilibrium continues to exist, the outcome of the game is unchanged if both the winner and the loser of the coin flip incur the development costs.

We are now prepared to define an equilibrium in this development game. Given the generality of our payoff structure, in which a firm may prefer losing sooner to losing later, it is important to disallow incredible threats where a firm attempts to bluff its rival into developing early. Such an equilibrium would occur when one firm announced a strategy of early development solely to induce its rival to preempt it (i.e., when the first firm would rather not develop at the early date it has announced). Our equilibrium concept, subgame perfect equilibrium, does not permit such incredible threats.

Formally, each firm's strategy indicates the dates at which it will choose to develop if development has not yet occurred. We restrict our attention to pure strategies. The notion of perfect equilibrium requires that the equilibrium strategies form a Nash equilibrium in every subgame; that is, that the players would choose to pursue their announced strategies in all possible contingencies (including those that do not actually arise in equilibrium). Thus, for example, firm i could not induce its rival to engage in early development by making a threat that firm i would not, in fact, wish to carry out.

B. *Characterization of Equilibrium*

Here we identify the basic incentives that determine the identity of the developing firm and the date of innovation. These incentives are expressed in terms of the reduced-form profit levels. In the following section, we relate these incentives to the underlying costs, imitation possibilities, and licensing fees.

Firm i's incentive to develop the innovation depends upon the profits that the firm would earn if it did *not* develop. These profits, in turn, depend on whether its rival, firm j, would develop the innovation if firm i did not.

First, suppose that firm j never would develop the innovation. If firm i chooses to develop at time T, its payoff is $W_i(T)$. We say that "firm i would develop on its own at T" if firm i would rather develop at T than have development never occur (i.e., if $W_i(T) \geq \pi_i^0/r$). Using equation (1), this inequality is equivalent to $\{\pi_i^i - \pi_i^0\}/r \geq K(T)e^{rT}$.

Since $K(T)e^{rT}$ declines with T, this condition either never will be satisfied (if $\pi_i^i - \pi_i^0 \leq rK$) or will be satisfied for all T at or after some finite time. In the case where firm i is willing to develop on its own at some dates, there is a unique date, $\hat{T}_i$, that maximizes $W_i(T)$. This date satisfies

$$(3) \qquad \pi_i^i - \pi_i^0 = -K'(\hat{T}_i)e^{r\hat{T}_i}.$$

If firm j never would develop, firm i would choose to develop at $\hat{T}_i$, which we call firm i's *stand-alone date* of development. The increment to flow profits earned by firm i when it innovates, $\pi_i^i - \pi_i^0$, is firm i's (flow) stand-alone incentive.

Now, suppose that firm i believes that if it did not develop the innovation at time T, firm j would. Then at time T, firm i faces the choice between developing and letting its rival win. We say that "firm i is willing to preempt at T" if $W_i(T) \geq L_i(T)$, i.e., if firm i would develop at T to avoid losing at T. Given our assumptions on costs, firm i's current value of preempting, $(W_i(T) - L_i(T))e^{rT}$, is increasing in T. Thus, if it is worth preempting at some date, it also will be worth preempting at any later date. Straightforward algebraic manipulations show that if $\pi_i^i - \pi_i^j > rK$, then there exists a unique date, $\tilde{T}_i$, such that firm i is willing to preempt at T if and only if $T \geq \tilde{T}_i$. $\tilde{T}_i$ is called firm i's *earliest preemption date*. If $\pi_i^i - \pi_i^j \leq rK$, then firm i never is willing to preempt.

When finite, $\tilde{T}_i$ is defined by $W_i(T) = L_i(T)$, or

$$(4) \qquad \pi_i^i - \pi_i^j = rK(\tilde{T}_i)e^{r\tilde{T}_i}.$$

The difference in profits from developing, rather than letting the rival develop, $\pi_i^i - \pi_i^j$, is firm i's (flow) preemption incentive. Observe that firm 1 has the greater preemption incentive if and only if post-innovation *industry* profits are higher following firm 1's innovating than they are subsequent to innovation by firm 2: the inequality $\pi_1^1 - \pi_1^2 \geq \pi_2^2 - \pi_2^1$ is equivalent to $\pi^1 \geq \pi^2$.

The $\hat{T}_i$'s and $\tilde{T}_i$'s are central to our analysis because they enter into some simple properties that all equilibrium outcomes must satisfy. By an outcome we mean the initial date of development and the identity of the winner. In the Appendix, we establish the following necessary conditions:

1) If firm i wins, it does so by the later of $\hat{T}_i$ and $\tilde{T}_j$.

2) If firm i wins, it does so at or after the earlier of $\hat{T}_i$ and $\tilde{T}_i$.

3) If firm i develops at any time other than $\hat{T}_i$, it must do so at $\tilde{T}_j$. When $\tilde{T}_i < \tilde{T}_j$, only firm i can win in this way.

4) If the firms tie, they must do so at $\tilde{T}_1 = \tilde{T}_2$. Therefore, ties are possible only if $\pi^1 = \pi^2$.

5) If $\hat{T}_i < \hat{T}_j$ and $\tilde{T}_i < \tilde{T}_j$, firm j cannot win.

We are now ready to characterize the equilibria. Adopt the labeling convention that post-innovation industry profits are at least as large when firm 1 innovates as they are when firm 2 does so: $\pi^1 \geq \pi^2$. Suppose for now that this inequality is strict. Putting the necessary conditions together, we have shown that if firm 1 wins, it does so either at $\hat{T}_1$ or at $\tilde{T}_2$. If firm 2 wins, it does so at $\hat{T}_2$, with $\hat{T}_2 < \hat{T}_1$. To determine which outcome prevails, we must compare the preemption and stand-alone incentives of the two firms. With our labeling convention, firm 1 has the greater preemption incentive. There are two cases to consider, depending upon whether firm 1 also has the greater stand-alone incentive.

First, consider the case where firm 1 has uniformly greater incentives to develop. Our first theorem shows that in this case, firm 1 wins, if anyone does.

THEOREM 1: *Suppose that firm 1 has greater preemption and stand-alone incentives (i.e., $\pi^1 > \pi^2$ and $\pi_1^1 - \pi_1^0 > \pi_2^2 - \pi_2^0$).*

(a) *If $\pi_1^1 - \pi_1^0 > rK$ and $\pi_1^1 - \pi_1^2 > rK$, then there exists a unique equilibrium outcome: firm 1 wins at the earlier of $\hat{T}_1$ and $\tilde{T}_2$.*[14]

[14] If $\pi_1^1 - \pi_1^0 > rK$ but $\pi_1^1 - \pi_1^2 \leq rK$, then $\tilde{T}_1 = \tilde{T}_2$ ($=\infty$) despite firm 1's greater preemption incentives. The general case in which $\tilde{T}_1 = \tilde{T}_2$ is discussed following Theorem 2.

(b) *If* $\pi_1^1 - \pi_1^0 \leq rK$, *then there exists an equilibrium without development. If, in addition,* $\pi_2^2 - \pi_2^1 > rK$, *then there exists a second equilibrium outcome: firm* 1 *wins at* $\tilde{T}_2 < \infty$. *In this case, both firms prefer the equilibrium involving no development, but firm* 1 *develops in "self-defense," for fear that firm* 2 *would otherwise do so.*

Formal proof of the theorems are available in our 1984 discussion paper.

Here we illustrate the various possible equilibria using figures. An example of the situation in which firm 1 wins at $\hat{T}_1 < \tilde{T}_2 < \infty$ is shown in Figure 1A. Firm 1's strategy calls for development on the interval $[\hat{T}_1, \infty)$, and firm 2's strategy calls for development on $[\tilde{T}_2, \infty)$. Firm 1's stand-alone incentive is strong enough that firm 2 is not willing to preempt firm 1 to keep from losing at $\hat{T}_1$. Firm 1 develops the innovation at the same date it would choose if it faced no *R&D* rivalry from firm 2. Figure 1B illustrates the standard patent race in which firm 1 preempts firm 2 by moving at $\tilde{T}_2$. Firm 1's strategy calls for development on $[\tilde{T}_2, \infty)$ and firm 2's for development on $(\tilde{T}_2, \infty)$. We reserve further discussion of these figures, and our interpretation of Theorems 1 and 2, for the following section.

Now, consider the case in which firm 2 has greater stand-alone incentives, and thus $\hat{T}_2 < \hat{T}_1$, while firm 1 continues to have a greater incentive to preempt, $\tilde{T}_1 < \tilde{T}_2$. In this "mixed" case, equilibria in which either firm wins may arise:

THEOREM 2: *Suppose firm* 1 *has greater preemption incentives but firm* 2 *has greater stand-alone incentives (i.e.,* $\pi^1 > \pi^2$ *and* $\pi_1^1 - \pi_1^0 < \pi_2^2 - \pi_2^0$*).*

(a) *If and only if* $\pi_2^2 - \pi_2^0 \leq rK$, *then there exists an equilibrium without development. If, in addition,* $\pi_2^2 - \pi_2^1 > rK$, *then there exists a second equilibrium outcome: firm* 1 *wins at* $\tilde{T}_2 < \infty$.

(b) *If* $\pi_2^2 - \pi_2^0 > rK$, *then any one of the following possibilities may arise:* (i) *there exists no equilibrium;* (ii) *firm* 1 *wins at the earlier of* $\hat{T}_1$ *and* $\tilde{T}_2$; *or* (iii) *firm* 2 *wins at* $\hat{T}_2$.

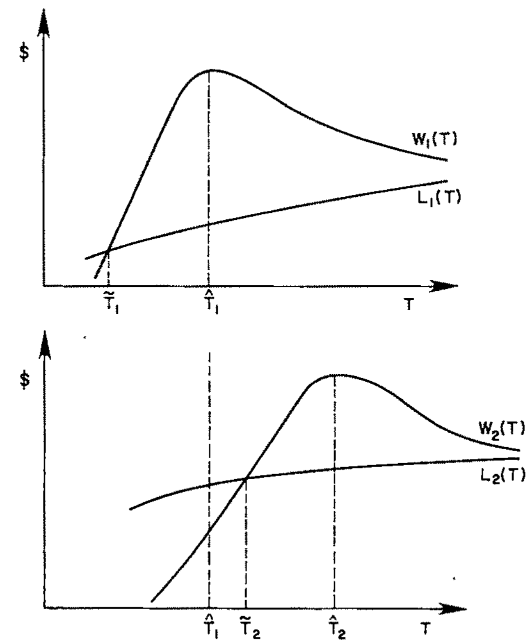


FIGURE 1A. FIRM 1 WINS AT $\hat{T}_1$

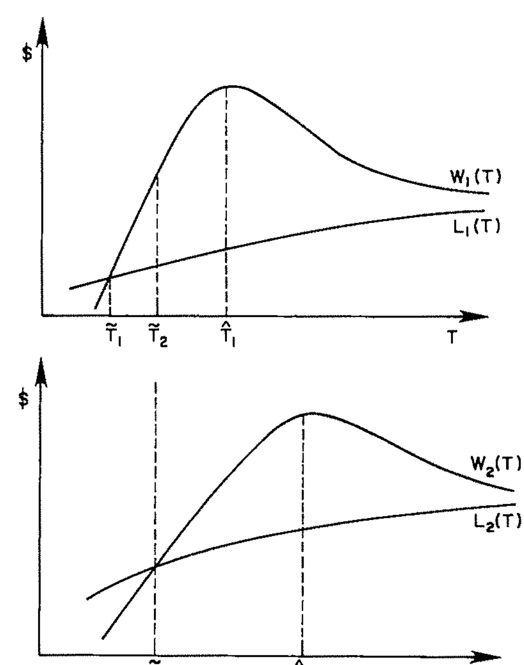


FIGURE 1B. FIRM 1 WINS AT $\tilde{T}_2$

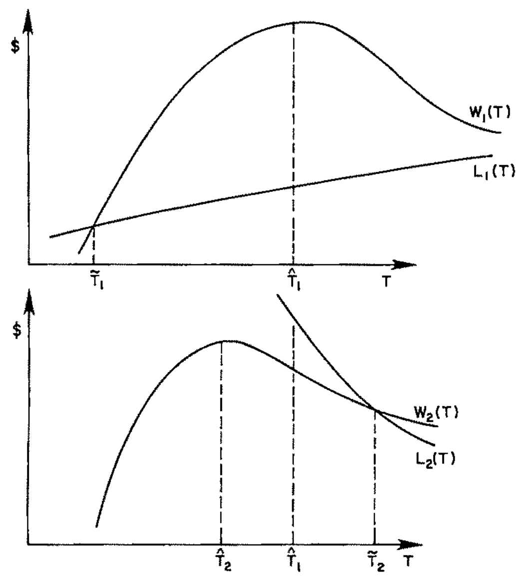


FIGURE 2A. FIRM 2 WAITS AND LOSES AT $\hat{T}_1$

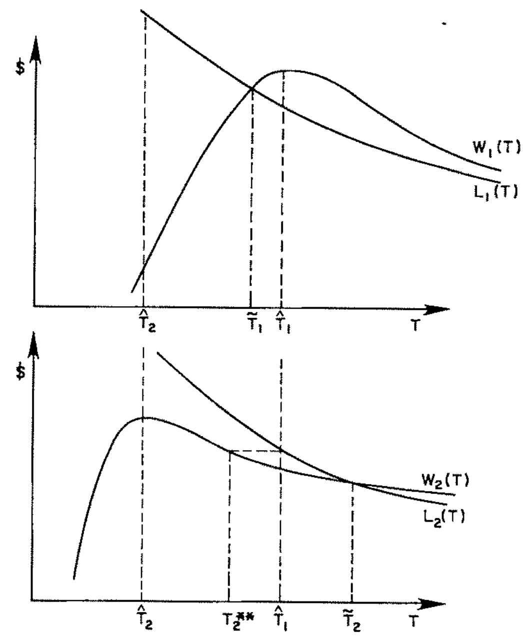


FIGURE 2B. FIRM 2 WINS AT $\hat{T}_2$

Figure 2A displays an equilibrium in which firm 1 wins at $\hat{T}_1$ despite firm 2's having stronger stand-alone development incentives. Firm 1's strategy calls for development on $[\hat{T}_1, \infty)$, and firm 2's for development on $[\tilde{T}_2, \infty)$. As drawn, firm 2 could win at $\hat{T}_2$, but prefers to wait and lose at $\hat{T}_1$. Knowing that firm 2 will wait until after $\hat{T}_1$ to develop, firm 1 waits until $\hat{T}_1$, its most-preferred development date.

Figure 2B shows payoff functions that are consistent with firm 2's winning at $\hat{T}_2$. Firm 2 would develop on $[\hat{T}_2, T_2^{**}]$, would not on $(T_2^{**}, \tilde{T}_2)$, and would resume at $\tilde{T}_2$. T_2^{**} is defined by $W_2(T_2^{**}) = L_2(\hat{T}_1)$. Firm 1 develops on $[\tilde{T}_1, \infty)$. With $\pi^1 > \pi^2$, this start/stop/start pattern by firm 2 must occur in any equilibrium in which firm 2 wins.

What features of the industry make it possible for firm 2 to win the development competition, despite the lower resulting industry profits? Surprisingly, a necessary condition for firm 2 to win is that firm 2 *benefit* when firm 1 develops (a possibility that arises only with licensing or imitation). If firm 2 is harmed by firm 1's developing the innovation, then firm 2's preemption incentive is stronger than its stand-alone incentive and $\tilde{T}_2 < \hat{T}_2$. But we know that firm 1 has greater preemption incentives than does firm 2, so $\tilde{T}_1 < \tilde{T}_2$ and firm 1 would preempt at any date on which firm 2 would want to move. This argument proves that firm 2 can develop the innovation only if $\pi_2^2 - \pi_2^0 > \pi_1^1 - \pi_1^0$ and $\pi_2^1 > \pi_2^0$.

Figure 2C illustrates the possible nonexistence of equilibrium. Firm 2 would prefer winning at $\hat{T}_2$ to losing at the earliest possible date that firm 1 would ever move: $W_2(\hat{T}_2) > L_2(\tilde{T}_1)$. While it is intuitively "obvious" that firm 2 must win at $\hat{T}_2$, formally there is no equilibrium in the continuation game beginning at $\tilde{T}_1$.[15] Thus, equilibrium fails to exist.

[15] In particular, there is no pure strategy equilibrium in the interval $[\tilde{T}_1, T_2^{**}]$. One solution to this "unconvincing" nonexistence problem is to resort to mixed strategies. Another is to revise our equilibrium concept along the following lines. So long as firm 1 avoids dominated strategies, firm 2 will wish to develop at $\hat{T}_2$, independent of firm 1's actual strategy subsequent to $\hat{T}_2$. In such a situation, we might declare the outcome "firm 2 wins at $\hat{T}_2$" to be the equilibrium outcome.

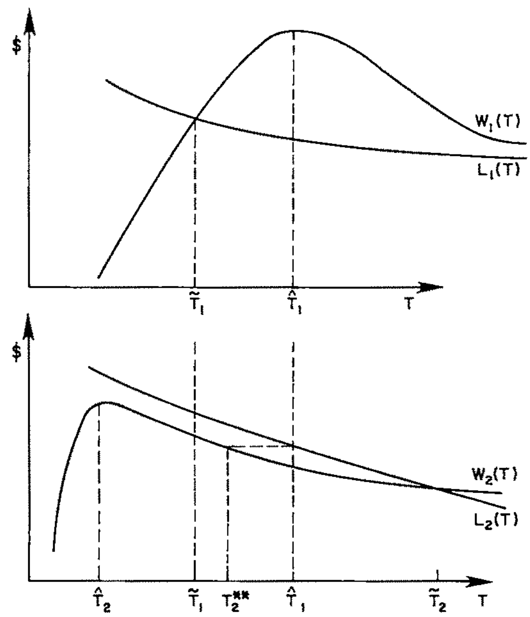


FIGURE 2C. NONEXISTENCE OF EQUILIBRIUM

Up to this point, we have assumed that post-innovation industry profits are strictly greater when one firm develops rather than the other. When post-innovation industry flow profits are independent of the identity of the winner, that is, $\pi^1 = \pi^2$ (or when both $\pi_1^1 - \pi_1^2$ and $\pi_2^2 - \pi_2^1$ are less than rK), $\tilde{T}_1 = \tilde{T}_2$. If $\tilde{T}_1 = \tilde{T}_2 < \min\{\hat{T}_1, \hat{T}_2\}$, then the firms tie at a development date of $\tilde{T}_1$. If $\hat{T}_i < \tilde{T}_i < \hat{T}_j$, then the unique equilibrium entails firm i's developing at $\hat{T}_i$. Lastly, if $\hat{T}_i < \hat{T}_j < \tilde{T}$ there exists one equilibrium under which firm i wins at $\hat{T}_i$; there exists a second equilibrium under which firm j wins at $\hat{T}_j$ if and only if $W_i(\hat{T}_i) \le L_i(\hat{T}_j)$.

II. Applications and Interpretation

In this section, we relate the equilibrium outcome to such basic factors as the firms' initial costs, the size of the innovation, the costs of imitation, and the bargaining institutions between the licensor and the licensee. Our examples all utilize the exponentially declining development cost function, $K(T) = K_0 e^{-\lambda T}$, with $\lambda > r$. This function obeys all of our development cost assumptions. For this $K(\cdot)$ function, the equations for $\hat{T}_i$ and $\tilde{T}_i$ are

$$(5) \qquad \hat{T}_i = \frac{1}{(\lambda - r)} \log \frac{\lambda K_0}{(\pi_i^i - \pi_i^0)};$$

$$(6) \qquad \tilde{T}_i = \frac{1}{(\lambda - r)} \log \frac{r K_0}{(\pi_i^i - \pi_i^j)}.$$

We employ a linear demand, Cournot duopoly structure in order to relate the flow profits to underlying industry conditions. The price received by the firms for their homogeneous product is given by $p = \alpha - x_1 - x_2$, where $\alpha > 0$ and x_i is firm i's output. The firms choose their outputs simultaneously, with each firm assuming that its rival's output is fixed. If firm i has constant marginal cost c_i, the Cournot equilibrium entails outputs of $x_i = (\alpha - 2c_i + c_j)/3$ and corresponding profits of $\pi_i = (\alpha - 2c_i + c_j)^2/9$, $i = 1, 2$, so long as each firm is active.

Until one of the firms chooses develop, firm i produces using the in-place technology at constant marginal cost c_i^0. Now label the firms so that, prior to either firm's innovating, firm 1 is the industry leader; $c_1^0 < c_2^0$. Firm 1's initial cost advantage is $c_2^0 - c_1^0$.

When firm i develops the innovation, its costs fall from c_i^0 to c_i^i. We denote firm 1's cost reduction by $\delta \equiv c_1^0 - c_1^1 \ge 0$. Define the *extra* cost reduction that firm 2 enjoys from the innovation by ε, so that firm 2's total cost reduction is $\delta + \varepsilon = c_2^0 - c_2^2$. We restrict attention to values of ε between $-\delta$ and $c_2^0 - c_1^0$. When $\varepsilon = -\delta$, $\delta + \varepsilon = 0$ and the innovation is of no value to firm 2. The case of $\varepsilon = c_2^0 - c_1^0$ can be thought of as one in which the innovation is sufficiently radical that it replaces the existing technology and allows either firm to produce at the same final cost ($c_1^1 = c_2^2$). In intermediate cases, the innovation is of value to both firms, and the leader's costs if it innovates are lower than the follower's costs would be if it innovated.

The noninnovating firm may also experience a cost change when development occurs. Firm j's costs fall from c_j^0 to c_j^i when firm i innovates; this cost reduction reflects firm j's ability to imitate firm i's innovation

or to lower its production costs via a licensing agreement with firm i.

A. *No Imitation or Licensing*

We consider first the pattern of innovation in the absence of licensing or imitation. This case arises when it is possible to protect the innovation through secrecy, but filing for a patent (in order to engage in licensing) would reveal information that allows rivals to "invent around" the patent.

Does firm 1 have sufficiently strong development incentives to remain the industry leader, or does firm 2 have stronger incentives to catch up or even take the lead? Without licensing or imitation, $c_j^i = c_j^0$ and $\pi_i^0 > \pi_i^j$. Consequently, each firm's preemption incentives are larger than its stand-alone incentives; assuming that $\tilde{T}_i < \infty$, $\tilde{T}_i < \hat{T}_i$ for both i. It follows that a necessary and sufficient condition for firm i to win is that it have stronger preemption incentives than firm j. Recall that firm i has greater preemption incentives than firm j if and only if industry profits are higher when firm i innovates than when firm j does so.

Gilbert and Newbery have examined the case of competition for an innovation where firm 1 would have a monopoly if it were to develop the innovation, but there would be a duopoly if firm 2 were to do so. In our notation, this case arises if $c_2^0 \geq (\alpha + c_1^1)/2$ and $c_1^0 < (\alpha + c_2^2)/2$. Given $c_1^1 \leq c_2^2$, the monopoly profits exceed the duopoly profits and firm 1 has the greater preemption incentives. Thus, the industry leader wins by developing at the earlier of $\hat{T}_1$ and $\tilde{T}_2$.[16] In their interpretation of this persistence of monopoly result, Gilbert and Newbery emphasize the case in which firm 1 has a monopoly prior to either firm's innovating, but the same analysis holds when there is a pre-innovation duopoly. As long as the innovation is essential for the follower's survival, the leader preempts. For $\delta = 0$, firm 1 will hold the innovation as a sleeping patent.

[16] In the related analysis of Reinganum (1983), uncertainty causes the preemption and stand-alone incentives to be mixed together. The greater stand-alone incentives of the industry follower then come into play, and that firm may win the development competition.

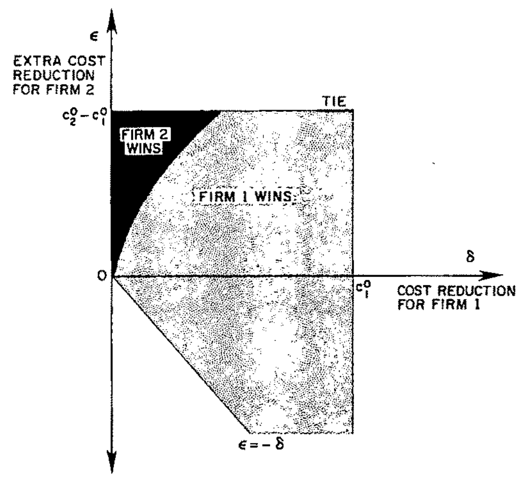


FIGURE 3. DEVELOPMENT WITH NO LICENSING OR IMITATION

Using our framework, we can generalize the persistence of industry leadership to cases in which there is duopoly both before and after development (i.e., $c_2^0 \leq (\alpha + c_1^1)/2$). Figure 3 shows how the two firms' preemption incentives depend upon the two crucial aspects of the innovation, δ and ε. The heavy line in Figure 3 divides the region in which the leader wins from that in which the follower wins. As the figure shows, the follower can win only if the innovation is of relatively small value to the leader (δ small) and yet of large value to the follower (ε large).

The cases considered here suggest two basic findings when *R&D* rivalry takes the form of a race. First, when the innovation is important to the survival of the industry follower, the industry leader will tend to innovate first in order to preempt and exclude its rival (even if the innovation is otherwise valueless to the leader). Second, the initial industry leader tends to be the innovator except when the innovation lowers the costs of the follower by significantly more than it would lower the costs of the leader.

B. *Imitation*

We now suppose that imitation is possible, but continue to assume that licensing is not. There are two dimensions to the ease of imitation. First, there is the degree to which imitation is feasible. For simplicity, we restrict our attention here to perfect imitation: $c_j^i = c_j^j$. Our results continue to hold for somewhat less-than-perfect imitation. Second, there is the cost of imitation. A simple and useful parameterization is to define γ_i as the ratio of the imitation costs to development costs for firm i. As noted in the introduction, Mansfield et al. (1981) found the average value of γ_i in their sample to be 0.65.

Given perfect imitation and $\gamma_i < 1$, imitation is more profitable than development. Hence, $\tilde{T}_1 = \infty = \tilde{T}_2$ and the development competition is a waiting game. In this waiting game, the firms' stand-alone incentives are critical to the development date and the identity of the winner. Firm i's stand-alone incentive equals $(\alpha - 2c_i^i + c_j^i)^2/9 - (\alpha - 2c_i^0 + c_j^0)^2/9$. Figure 4 indicates the regions in (δ, ε) space in which each firm has the greater stand-alone incentive.

If firm 1's stand-alone incentive is small relative to that of firm 2, then the unique equilibrium entails firm 2's winning at $\hat{T}_2$. In this case, the industry follower is too impatient to wait for development by the leader. This certainly will be the outcome if $\varepsilon \geq \delta$, in which case $\pi_1^1 \leq \pi_1^0$ and firm 1 never will develop on its own. Likewise, if firm 2's stand-alone incentives are relatively small, the unique equilibrium is for firm 1 to win at $\hat{T}_1$. When the stand-alone incentives are roughly equal, however, each of these outcomes is an equilibrium.[17] Therefore, near the boundary in Figure 4, either firm can be the innovator.

Our theory generates a number of predictions for cases in which imitation is easy. First, if $\varepsilon \leq 0$, then the industry leader has

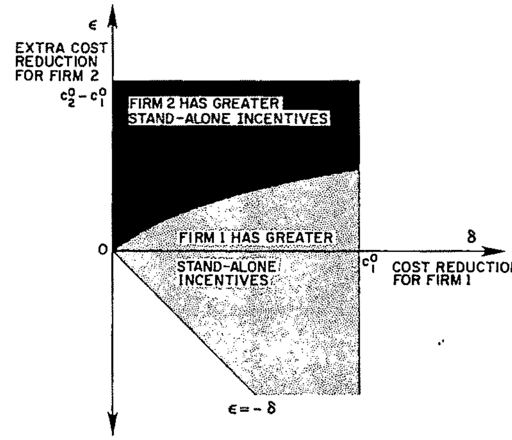


FIGURE 4. DEVELOPMENT WITH IMITATION

the greater stand-alone incentives and there exists an equilibrium in which the leader innovates at $\hat{T}_1$. Of course, there also may exist an equilibrium in which the follower innovates. Multiple equilibria also arise when the leader's initial advantage $(c_2^0 - c_1^0)$ is small and the innovation is of approximately equal value to each firm (ε near zero). Finally, and most importantly, our theory predicts that when one firm has a substantial technological lead ($c_2^0 - c_1^0$ large), that firm will be reluctant to make innovations that are easily imitated and tend to equalize costs *across* the industry (i.e., ε near $c_2^0 - c_1^0$). Instead, such innovations will come from firms that have high costs and low market shares prior to the innovation.

C. *Patent Licensing*

Now, suppose that licensing is feasible. For simplicity, we focus our attention on the case in which licensing permits perfect transfer of technological know-how, so that firm j obtains the same cost reduction as licensee or innovator. We suppose also in our licensing examples that imitation is so costly that in the absence of licensing the noninnovator would not obtain the innovation.

In addition to the production cost changes induced by the innovation, the firms' post-innovation profits depend upon licensing fees

[17]If firm i has the greater stand-alone incentive, the condition for multiple equilibria to exist is $W_i(\hat{T}_i) \leq L_i(\hat{T}_j)$.

and technology transfer costs. Technology transfer costs are borne by the losing firm, but not received by the winner. We denote by M_j the exogenously given capital equipment and training costs incurred by firm j after firm i develops. On a flow basis, the one-time cost M_j is amortized as rM_j. In our examples, we assume that $M_1 = M_2$.[18] Firm j also pays a flow licensing fee, L_j^i to firm i for the use of the new technology.[19] Unlike the technology transfer cost, the licensing fee constitutes an addition to the licensor's profits. The license fee also differs from the transfer cost in that the fee is endogenous.

The license fee is determined through negotiations between the two firms. The threat points in these negotiations are the profits that would arise in the absence of licensing. Suppose that firm i innovates. The no-licensing profits are $(\alpha - 2c_i^i + c_j^0)^2/9$ for firm i and $(\alpha - 2c_j^0 + c_i^i)^2/9$ for firm j. With licensing, firm i earns $\pi_i^i = (\alpha - 2c_i^i + c_j^i)^2/9 + L_j^i$ while firm j earns $\pi_j^i = (\alpha - 2c_j^i + c_i^i)^2/9 - rM_j - L_j^i$. Comparing its two profit levels, the lowest fee that firm i would accept is

$$(7)\quad \underline{L}_j^i = \left(c_j^0 - c_j^i\right)\left\{2\alpha - 4c_i^i + c_j^0 + c_j^i\right\}/9.$$

Likewise, the highest fee that firm j would offer is

$$(8)\quad \overline{L}_j^i = 4\left(c_j^0 - c_j^i\right) \times \left\{\alpha + c_i^i - c_j^0 - c_j^i\right\}/9 - rM_j.$$

We assume that the two firms will strike a licensing deal if and only if there are licensing gains from trade (i.e., $\overline{L}_j^i - \underline{L}_j^i > 0$). Depending upon the bargaining institutions, the actual license fee may fall anywhere in the range from $\underline{L}_j^i$ to $\overline{L}_j^i$. Let σ denote the licensor's share of the gains from trade, $0 \leq \sigma \leq 1$. Extreme values of σ correspond to take-it-or-leave-it offers by one side or the other. The Nash Bargaining solution yields a value of $\sigma = \frac{1}{2}$.

Figure 5A shows the regions of (δ, ε) space in which firms 1 and 2 would license the innovation.[20] As the figure indicates, there are three possibilities: 1) neither firm would license; 2) only the follower, firm 2, would license; or 3) each firm would license. We discuss these cases in turn.

Case 1: *Neither Firm Would License*. When the innovation leads to large cost reductions, a firm may refuse to license to its rival. For example, any innovation sufficiently large to grant the innovator a monopoly falls into this category. We know from Part A above that, in the absence of licensing, the preemption incentives determine the identity of the winner. Simple calculations show that for almost all of the innovations in the no-licensing region of Figure 5A, firm 1 has the higher preemption incentive (compare with Figure 3). The only exceptions are innovations along the line $\varepsilon = c_2^0 - c_1^0$ that are so drastic that they would allow either firm a monopoly; for these innovations, the two firms' preemption incentives are equal. We conclude that the industry leader typically develops innovations that would not be licensed.

Case 2: *Only Firm 2 Would License*. In some cases, only one of the two firms would license. Direct calculation shows that the initial industry leader must be the excluding firm. Since firm 1 would not license and there is no imitation, firm 2 cannot benefit from firm 1's development of the innovation. Therefore, firm 2's preemption incentive exceeds

[18] When the technology transfer cost varies across firms, the firm to whom it is more costly to transfer technology has an additional incentive to develop the innovation.

[19] We restrict attention to license fees that are independent of the output of the licensee. For analyses of per-unit license fees, see Kamien and Tauman, our 1985 paper, and Shapiro (1985).

[20] In our 1985 paper, we derive more general conditions under which direct rivals choose to license innovations in markets where imitation is impossible. Of course, licensing still may take place when patents and secrecy are ineffective: a licensing arrangement can amount to an admission that, absent licensing, imitation would occur.

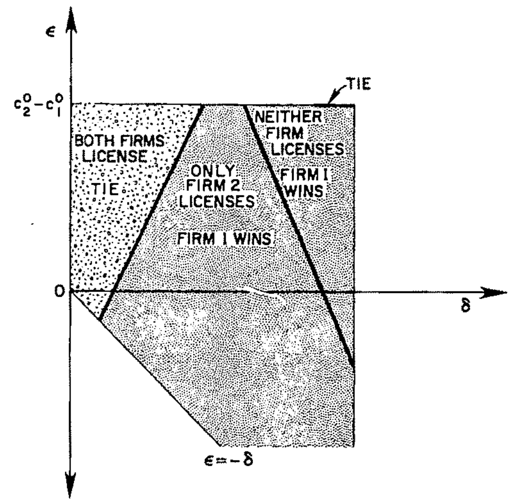


FIGURE 5A. DEVELOPMENT WITH PATENT LICENSING ($\sigma = 1$)

Note: For larger values of $c_2^0 - c_1^0$, only firm 2 would license innovations near the origin.

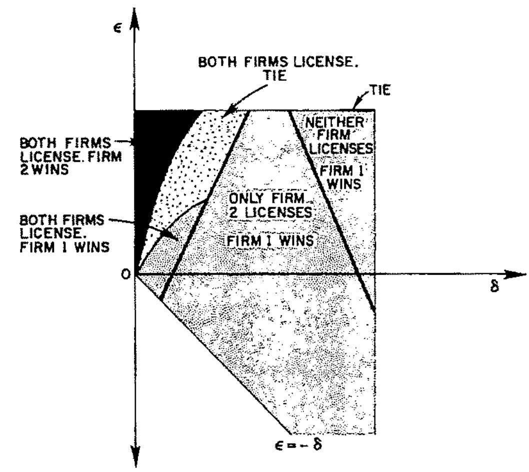


FIGURE 5B. DEVELOPMENT WITH PATENT LICENSING ($\sigma = 0$)

its stand-alone incentive, and $\tilde{T}_2 < \hat{T}_2$. Since industry profits are higher when firm 1 innovates than when firm 2 does so, firm 1's preemption incentive exceeds firm 2's, so $\tilde{T}_1 < \tilde{T}_2$.[21] These two inequalities imply that the industry leader must develop the innovation.

Case 3: *Each Firm Would License*. If each firm would license, the two firms have equal preemption incentives. For values of σ close to 1, the gains from licensing are appropriated by the licensor and the licensee cannot benefit from development. Hence, for large values of σ, the preemption incentives determine the identity of the winner ($\tilde{T}_i < \hat{T}_i$), and either firm can develop the innovation. Figure 5A illustrates this case. The follower can win even when it values the innovation less for its own use than does the leader, because the follower earns a larger licensing fee.

For lower values of σ, the stand-alone incentives may come into play. Suppose, for example, that $\sigma = 0$. A value of $\sigma = 0$ does not mean that the innovator gives away its technology. Rather, it corresponds to a situation in which the license fee just compensates the innovator for having to face a more efficient rival. Figure 5B illustrates the dependence of the equilibrium outcome on the values of δ and ε when $\sigma = 0$.

Although they differ in details, the overall stories told by Figures 5A and 5B are similar: when fixed fee licensing is feasible, the industry leader tends to be the innovator except for innovations such that δ is small and ε is large. This pattern is similar to the one that arises in the absence of imitation or licensing.

D. *Interpretation and Empirical Support*

Our theory generates predictions for the pattern of development of innovations by industry leaders and followers. The theory predicts that, with or without licensing or imitation, the industry leader will tend to be the innovator unless the innovation is such that the follower gets a significantly larger

[21] Since firm 1 chooses not to license, licensing must lower industry profits. But licensing by firm 1 would establish the same level of profits earned when firm 2 innovates, namely the level that applies when each firm can use the innovation. Therefore, $\pi^1 > \pi^2$ when only firm 2 would license.

cost reduction than does the leader.[22] Thus, for innovations that reduce costs nearly equally (ε/δ small), we expect industry leaders to be innovators. For innovations such that both δ and ε are large, the identity of the winner depends on whether imitation is possible. Absent imitation, the leader will tend to develop such innovations even when the follower would benefit from a greater cost reduction than would the leader. When imitation is easy, however, industry followers or entrants will tend to develop such innovations.

In relating our predictions to existing evidence, it is useful to think in terms of major and minor innovations. We define a *minor innovation* as one for which δ is small and for which ε/δ is small. In contrast, a *major innovation* is one for which δ is large relative to c_1^0 and ε is close to $c_2^0 - c_1^0$. Minor innovations are in the southwest corner of the positive quadrant in Figures 3, 4, 5A, and 5B, while major innovations are in the northeast. In terms of these definitions, our two central predictions are that: 1) the leader will tend to develop minor innovations, with or without imitation or licensing; and 2) the leader will tend to develop major innovations if and only if imitation is difficult.

We are not aware of existing empirical studies using our definitions of major and minor innovations. In particular, the literature does not indicate whether a given innovation reduced one firm's costs by more than it would have reduced others'. The commonly used definitions of major and minor innovations focus on the absolute size of the cost reduction enjoyed by the innovator. We believe, however, that when the innovation is a major one as measured by $(c_i^0 - c_i^i)/c_i^0$, it is likely to entail the development of an entirely new process or product. Consequently, the post-innovation cost level may well be insensitive to the firm's cost level under the old technology. Thus, we hypothesize that such innovations will tend to exhibit greater cost reductions for initially high-cost firms (i.e., ε will be near $c_2^0 - c_1^0$). On the other hand, when the innovation is an incremental improvement in some component of the existing production process, it is unlikely to narrow the gap between the leader and the follower. In fact, the extent of cost reduction may be independent of the firms' initial costs. Thus, we believe that, in practice, the labeling of innovations as major and minor under our definitions will tend to conform with that of the empirical literature.

We are unaware of any direct evidence bearing on our prediction regarding minor innovations.[23] However, our second prediction, that major innovations will be developed by industry leaders if and only if imitation is difficult, is supported by existing empirical studies. Mansfield (1981) provides the only cross-industry survey of which we are aware on whether small or large firms tend to develop major innovations. Four of the twelve industries that he examined, chemicals, automobiles, petroleum, and office equipment and computers, had elasticities of expenditure on the development of entirely new products and processes with respect to firm sales that were greater than unity. Drugs were the only other industry with an elasticity above the mean. Industry studies corroborate the view that large firms were responsible for most of the major innovations in the chemical, petroleum, and drug industries.[24]

These data are largely consistent with our prediction that this pattern of relatively intensive $R\&D$ by established firms should be most pronounced in industries where imitation is relatively difficult. The chemical, petroleum, and drug industries often are cited as the ones in which imitation is most dif-

[22]When ε/δ is small, the follower can win, but only by tying the leader (if both firms license and σ is large) or due to the presence of multiple equilibria (which allow the follower to win despite its smaller stand-alone incentives). It is in this sense that we say that the industry leader tends to win when ε/δ is small.

[23]There is some indirect evidence suggesting that industry leaders introduce a disproportionate number of minor innovations. As Scherer discusses, leaders spend more on $R\&D$ and produce more patents than do followers. In industries where leaders are not responsible for a disproportionate share of major innovations, they must file for more minor patents than do followers.

[24]See, for example, Christopher Freeman (1974), Mansfield et al. (1977), and the studies that they cite.

ficult.[25] Office equipment and computers present something of a puzzle. Levin et al. (pp. 30; 34) indicate that in the computer industry imitation is relatively easy. Thus, our theory predicts that the leading firm should be an imitator. This prediction runs counter to Mansfield's finding for office equipment and computers as a whole. Gerald Brock (1975, pp. 185–207) and Christopher Freeman (1974, pp. 132–33), however, cite data indicating that the leading computer firm, IBM, tended to be an imitator rather than an innovator for major developments.[26] Hence, we believe that the computer portion of this industry fits our predicted pattern. Finally, we could not find data on the ease of imitation in the automobile industry.

We turn now to industries that Mansfield's survey identified as ones in which large firms did disproportionately little *R&D* aimed at entirely new products and processes. Our theory again is borne out in that imitation is both common and effective in these industries. In the machinery industry, Mansfield ranked the relative performance of large firms as eleventh out of the twelve industries surveyed. Electronics and electrical equipment ranked sixth out of twelve, while instruments ranked eighth. Mansfield et al. (1981) found that imitation was relatively easy in the combined area of electronics and machinery. Levin et al. (p. 20) found that process innovations were easily imitated in the machinery sector.[27] The food industry too, appears to be one in which imitation is easy (Levin et al., pp. 20; 33) and large firms tend to imitate rather than innovate (food was seventh on Mansfield's list).[28]

III. Imitation, Licensing, and Public Policy

Having characterized the structure of the equilibria, we turn now to effects of licensing or imitation on the date of development. These comparative statics are an important component of policy analysis. The goal of assigning intellectual property rights such as patents, copyrights, and trademarks to innovators is to give private agents incentives to engage in costly research and development activities. Does strict enforcement of a strong system of intellectual property rights in fact speed up the pace of innovation?

A. *The Effects of Imitation*

For simplicity, consider a market where licensing is blocked and the only possible means of dissemination of the innovation is through imitation. Both the extent to which imitation is imperfect and the cost of imitation affect the date of development. Consider first changes in firm i's costs of imitation as captured by the imitation efficiency parameter γ_i. Such changes could be due to a shift in patent policy, for example.

Suppose that imitation costs initially are low enough that imitation is profitable. An increase in firm i's imitation cost lowers π_i^j, but has no effect on the other flow profit levels as long as the cost change does not affect the outcome of the imitation decision. Thus, an increase in γ_i leaves $\tilde{T}_j$, $\hat{T}_1$, and $\hat{T}_2$ unaffected. $\tilde{T}_i$ is lower, as firm i is willing to fight harder to avoid its losing and having to engage in imitation.

[25] See, for example, Freeman, Levin et al., and Mansfield et al. (1981). We can distinguish between product and process innovations in the chemical industry. DuPont, the industry leader, played a much stronger role in major product innovations (Mansfield et al., 1977, p. 67) where patent protection is strong (Levin et al., p. 30), than in major process innovations (Mansfield et al., 1977, p. 66) where imitation was comparatively easy (Levin et al., p. 20).

[26] In addition to the difference in industry definitions, this discrepancy may come from the fact that Mansfield measured *R&D* expenditures, while Brock and Freeman looked at the actual pattern of development. The studies also considered different time periods.

[27] Levin et al. (p. 16) did find, however, that patent protection was relatively effective for product innovations in this sector.

[28] We were unable to obtain data on the ease of imitation for the remaining three industries surveyed in Mansfield: metals, aerospace, and soap and cosmetics. The steel industry, however, appears to be one in which imitation is easy (Levin et al., p. 20) and industry leaders are not the major innovators (Mansfield et al., 1977, p. 15).

How do these changes translate into shifts in the development date? Restricting attention to strict inequality of industry profits and labeling the firms so that $\pi^1 > \pi^2$, we have seen that there are three types of equilibrium to consider when calculating comparative statics: 1) firm 1 may win at $\hat{T}_1$; 2) firm 2 may win at $\hat{T}_2$; or 3) firm 1 may win by preempting firm 2 at $\tilde{T}_2$. We suppose, initially, that the changes in the underlying parameters do not alter the type of equilibrium.

Changes in γ_1 affect only $\tilde{T}_1$ and thus have no effect on the equilibrium outcome as long as the ordering between $\tilde{T}_1$ and $\hat{T}_2$ is unaffected. When γ_2 changes, the only effects on development arise in equilibria where firm 1 wins at $\tilde{T}_2$. Here, parameter changes have an *indirect* effect on the development date—the change in the development date is brought about by the change in the willingness of the *losing* firm (firm 2) to engage in preemption. The winning firm (firm 1) moves just soon enough to keep its rival from engaging in preemptive development. Raising the cost of imitation leads to earlier development, but not because it has made moving first more profitable. Rather, firm 1 develops sooner because it has to fight harder to keep from losing.

A second change in the imitation technology would be a shift in the cost reduction enjoyed by the imitating firm, as captured by $c_j^i - c_j^j$.[29] A decrease in firm j's ability to copy from firm i would lower π_j^i (since firm j would have higher costs subsequent to i's innovation) and raise π_i^i (since the innovator would retain a larger cost advantage). In this case, both the direct and indirect effects would lead to earlier development in any of the three types of equilibrium.

We also must consider the possibility of regime changes due to shifts in the imitation technology. When the regime change is between type 1 and type 3 equilibria, the initial development date moves continuously with $\hat{T}_i$ and $\tilde{T}_i$; the change in regime occurs only when $\hat{T}_1 = \tilde{T}_2$. Thus, any parameter change that increases both the preemption and the stand-alone incentives leads to earlier innovation in this case. The more interesting regime changes are those involving the type 2 equilibrium in which firm 2 wins at $\hat{T}_2$. Starting at such an equilibrium, a small change in the underlying parameters may lead to either nonexistence of equilibrium, or a move to a type 1 equilibrium. Such changes can have surprising effects, as the following example illustrates.

When the move is from a type 2 to a type 1 equilibrium, the date of development jumps from the original value of $\hat{T}_2$ to the new value of $\hat{T}_1$. In the original equilibrium, $\hat{T}_2 < \hat{T}_1$, and this relationship may continue to hold after the change of regime. Thus, a parameter shift that moves both $\hat{T}_i$ and $\tilde{T}_i$ forward in time for $i = 1,2$ may nevertheless lead to later development. For example, *reduced imitation* possibilities may *slow down* initial development. Firm 1, knowing that imitation is less complete, is more eager to develop and does so sooner. As a consequence, firm 2 decides to wait and imitate rather than to develop on its own.

B. *The Effects of Licensing*

We turn now to the effects of changes in the licensing "technology." Suppose that the costs of transferring innovative know-how to firm j through a license, M_j, fall. Conditional on licensing taking place, a decline in these costs results in an increase in post-development industry profits when firm i is the innovator. If the licensor's share of the gains is σ, then π_i^i rises by σ and π_j^i rises by $1-\sigma$ in response to a unit reduction in licensing costs. In the natural case where both M_1 and M_2 fall equally, $\hat{T}_1$ and $\hat{T}_2$ both fall as well since there are greater gains from development compared to no development. The effects on $\tilde{T}_1$ and $\tilde{T}_2$, however, depend upon the specific value of σ. $\tilde{T}_i$ comes sooner as licensing costs fall if and only if the $\sigma > 1/2$.

[29]An increase in the lag time needed to copy the innovation has the same qualitative effects as does an increase in the post-innovation production costs of the imitator.

Consider first the effects of changes in technology transfer costs on an equilibrium where firm i wins at $\hat{T}_i$. In this case, a reduction in technology transfer costs has a direct effect on the development date by raising firm i's stand-alone incentive and lowering $\hat{T}_i$; for all $\sigma > 0$ the direct effect of a reduction in licensing costs leads to earlier development, as we would expect.

If the equilibrium entails firm 1 winning at $\tilde{T}_2$, then a reduction in licensing costs leads to *less* rapid development if and only if $\sigma < 1/2$. If $\sigma < 1/2$, a decrease in licensing costs helps the loser more than the winner, and each firm is willing to wait longer before it will attempt to edge out its rival to prevent losing. The theory of bargaining predicts that the equilibrium value of σ is very sensitive to the structure of the bargaining institutions. In their empirical examination of licensing, Richard Caves et al. (1983) estimated that the licensor gains between 1/3 and 1/2 of the expected rents, with an average of approximately 40 percent. These data indicate that $\sigma < 1/2$ and hence that lower licensing costs slow development by making losing less painful.

We also must consider the possibility that the change in the underlying parameters alters the type of equilibrium that obtains. For example, suppose that there is a fall in the costs of licensing and that most of the gains are appropriated by the licensor. Then both $\hat{T}_1$ and $\hat{T}_2$ fall. Firm 2 now may prefer to wait and lose at the relatively early date $\hat{T}_1$ rather than to develop at $\hat{T}_2$. Firm 1's increased development incentives may actually delay innovation: the date of development may shift from the original value of $\hat{T}_2$ to the new, possibly later time of $\hat{T}_1$.

Patent policy can affect the licensing outcome in a variety of ways. For example, a weak patent policy (i.e., one that makes imitation less costly) may lower the bargaining power of the licensor by raising the licensee's profits at its threat point (imitation) in the negotiations. Our results show that this reduction in bargaining power typically would delay the initial date of development both by making winning less desirable and making losing less undesirable. A weak patent system could, however, induce earlier innovation by firm 2. Again, firm 1's reduced incentive to innovate raises $\hat{T}_1$ and may compel firm 2 to develop at an earlier date, $\hat{T}_2$.

IV. Conclusion

In this paper, we have constructed a framework for the analysis of innovation that goes beyond the existing literature by incorporating a payoff structure that is general enough to allow for the possibilities of licensing and imitation. Both of these extensions are important ones if this line of research is to serve as a basis for understanding actual *R&D* rivalry. The "patent race" may actually be a waiting game. With our more general payoff structure, we find that a firm may win easily, rather than just edging out its rival as in earlier models. In such cases, the existence of rivalry for product development has no effect on the date of development. Moreover, we show that a firm may win the deterministic patent race even though industry profits would be greater were the other firm to win.[30]

Our analysis generates some specific predictions about the pattern of innovation in an industry composed of small and large firms. For minor innovations, we find that the industry leader typically will be the innovator, whether or not imitation and licensing are feasible. For markets in which patent protection is strong, our theory predicts that the major innovations will be made by industry leaders. But if imitation is easy, industry followers or entrants will make the major discoveries. Preliminary analysis of

[30] We have not said anything about welfare. As in the rest of the patenting literature, there are several distortions present. Each firm cares only about its own profits. The firm ignores the effects that its actions have on both the profits of its rival and the welfare of consumers. Typically, these two wedges go in opposite directions (i.e., the rival's profits fall and consumers' surplus rises when a given firm patents), and either effect may dominate the other. See our 1987 paper for a more general treatment. In those cases where the losing firm and consumers both gain from the innovation, we know that a firm that wins easily has waited too long. Even in this case, however, we do not know the direction of the distortion when a firm preempts its rival.

available data suggests that this relationship between innovation and imitation holds in a variety of industries. Whether more detailed empirical work will reach the same conclusion remains to be seen.

We also find that changes in the licensing and imitation technologies, or in public policy treatment of licensing and imitation, may have surprising effects. A decrease in the transactions cost of licensing that raises the innovator's profits may lead to slower development by reducing the incentives of the losing firm to fight for the initial property rights. The division of the gains from licensing between the two firms is the essential parameter that determines whether reduced licensing costs leads to faster or slower development. As one would expect, we also find that a reduction in the costs of imitation will tend to delay development. But this effect arises through a reduction in the incentives of the losing firm: as imitation becomes cheaper, being the "loser" becomes more desirable. The incentives of the winning firm are unaffected by the change in imitation costs. Finally, we demonstrate that an increase in the incentives of each firm both to develop on its own and to preempt its rival still may lead to slower development.

APPENDIX

Proofs of Conditions Necessary for Equilibrium:

1. If firm i wins after $\hat{T}_i$, then it could have developed earlier and won at its favorite time, $\hat{T}_i$. Of course, for firm i to win, firm j must let it. Thus, firm i can win no later than $\tilde{T}_j$.

2. Suppose to the contrary that firm i wins at T_i^* which is before the minimum of $\hat{T}_i$ and $\tilde{T}_i$. If firm j's strategy calls for development at or just after T_i^*, then firm i should not move at that date; given $T_i^* < \tilde{T}_i$, it is better for firm i to let firm j win. So firm j's strategy must not call for development at or just after T_i^*. But then firm i could win a little later than at T_i^*, and it would prefer to do so since $W_i(T)$ is quasi concave and firm i would rather wait until $\hat{T}_i$ to win.

3. The only reason for firm i to win at some time other than its most-preferred one (i.e., $\hat{T}_i$), is to stop firm j from winning. Firm i will develop as soon as $\tilde{T}_i$ to preempt firm j, but firm j will not attempt to preempt before $\tilde{T}_j$.

4. We know that the firms cannot tie sooner than $\tilde{T}_i$, or else firm i would do better to let firm j win outright. If $\tilde{T}_1 \neq \tilde{T}_2$, we can label the firms such that $\tilde{T}_1 < \tilde{T}_2$, and firm 1 would break the tie by moving sooner than would firm 2.

5. Label the firms such that $\tilde{T}_1 < \tilde{T}_2$. We know that firm 2 can never develop before $\min(\hat{T}_2, \tilde{T}_2)$. If $\tilde{T}_2 < \hat{T}_2$, then $\tilde{T}_1 < \tilde{T}_2 < \hat{T}_2$, and firm 1 would preempt at any date on which firm 2 would develop. Suppose that $\hat{T}_2 < \tilde{T}_2$, and hence $\hat{T}_1 < \hat{T}_2 < \tilde{T}_2$. By Lemma 1 below, firm 2's strategy cannot call for development on $[\hat{T}_1, \tilde{T}_2)$, so it cannot call for development on $[\hat{T}_2, \tilde{T}_2)$. Therefore, firm 2 cannot develop at all before $\tilde{T}_2$. Again, firm 2 cannot win; firm 1 always would preempt it.

LEMMA 1: *If* $\hat{T}_1 < \tilde{T}_2$, *and* $\tilde{T}_1 < \infty$, *then firm 2's equilibrium strategy cannot call for development during the interval* $[\hat{T}_1, \tilde{T}_2)$.

PROOF:

It is a dominant strategy for firm 1 to develop at all $T \geq \max(\hat{T}_1, \tilde{T}_1)$. By definition, firm 2 will not develop at any date before $\tilde{T}_2$ on which firm 1 is. If $\tilde{T}_1 \leq \hat{T}_1$, then firm 1 develops on $[\hat{T}_1, \infty)$ and thus on $[\hat{T}_1, \tilde{T}_2)$, so firm 2 will not.

If $\tilde{T}_1 > \hat{T}_1$, then firm 1 still develops on $[\tilde{T}_1, \tilde{T}_2]$, so firm 2 will not. Therefore, if firm 2's strategy calls for him to develop on $[\hat{T}_1, \tilde{T}_2)$, he must stop strictly before $\tilde{T}_1$. Call T_2^* the *last* date on $[\hat{T}_1, \tilde{T}_2)$ at which firm 2's strategy calls for development. Since $T_2^* > \hat{T}_1$, firm 1's strategy must then call for development at $T_2^* + \Delta$; he would rather win then than later. But firm 2 would rather lose at $T_2^* + \Delta$ (for Δ small) than win at T_2^*, since $T_2^* < \tilde{T}_2$. This contradicts the existence of T_2^*, and proves that firm 2's strategy cannot call for development at all on $[\hat{T}_1, \tilde{T}_2)$.

REFERENCES

Brock, Gerald W., *The U.S. Computer Industry*, Cambridge: Ballinger, 1975.

Caves, Richard E., Crookell, Harold and Killing, J. Peter, "The Imperfect Market for Technology Licenses," *Oxford Bulletin of Economics and Statistics*, August 1983, *45*, 249–67.

Dasgupta, Partha and Stiglitz, Joseph E., "Uncertainty, Industrial Structure, and the Speed of *R&D*," *Bell Journal of Economics*, Spring 1980, *11*, 1–8.

Freeman, Christopher, *The Economics of Industrial Innovation*, Harmondsworth: Penguin Books, 1974.

Fudenberg, Drew and Tirole, Jean, "Preemption and Rent Equalization in the Adoption of New Technology," *Review of Economic Studies*, July 1985, *52*, 383–401.

______ **and** ______, "A Theory of Exit in Duopoly," *Econometrica*, July 1986, *54*, 943–60.

Gallini, Nancy T., "Deterrence by Market Sharing," *American Economic Review*, December 1984, *74*, 931–41.

______ **and Winter, Ralph,** "Licensing in the Theory of Innovation," *Rand Journal of Economics*, Summer 1985, *16*, 237–52.

Gilbert, Richard and Newbery, David, "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

Kamien, Morton and Tauman, Yair, "The Private Value of a Patent: A Game-Theoretic Analysis," *Quarterly Journal of Economics*, August 1986, *101*, 471–91.

Katz, Michael L., "An Analysis of Cooperative Research and Development," *Rand Journal of Economics*, Winter 1986, *17*, 527–43.

______ **and Shapiro, Carl,** "Perfect Equilibrium in a Development Game with Licensing or Imitation," Woodrow Wilson School Discussion Paper No. 85, Princeton University, December 1984.

______ **and** ______, "On the Licensing of Innovations," *Rand Journal of Economics*, Winter 1985, *16*, 504–20.

______ **and** ______, "How to License Intangible Property," *Quarterly Journal of Economics*, August 1986, *101*, 567–89.

______ **and** ______, "The Two Wedges: The Fundamental Theorem of Industrial Organization," Princeton University, 1987.

Levin, Richard C. et al., "Survey Research on *R&D* Appropriability and Technological Opportunity, Part I: Appropriability," Yale University, 1986.

Mansfield, Edwin, "Composition of *R&D* Expenditures: Relationship to Size of Firm, Concentration, and Innovative Output," *Review of Economics and Statistics*, November 1981, *63*, 610–15.

______, **Schwartz, Mark and Wagner, Samuel,** "Imitation Costs and Patents: An Empirical Study," *Economic Journal*, December 1981, *91*, 907–18.

______ **et al.,** *The Production and Application of New Industrial Technology*, New York: W. W. Norton, 1977.

Nalebuff, Barry and Riley, John, "Asymmetric Equilibria in the War of Attrition," *Journal of Theoretical Biology*, July 1984, *113*, 517–27.

Reinganum, Jennifer, (1981a) "On the Diffusion of New Technology: A Game-Theoretic Approach," *Review of Economic Studies*, July 1981, *48*, 395–405.

______, (1981b) "Dynamic Games of Innovation," *Journal of Economic Theory*, August 1981, *25*, 21–41.

______, "A Dynamic Game of *R&D*: Patent Protection and Competitive Behavior," *Econometrica*, May 1982, *48*, 671–88.

______, "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, September 1983, *73*, 741–48.

Rostoker, Michael D., "A Survey of Corporate Licensing," *IDEA*, 1984, 24.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, Chicago: Rand McNally, 1980.

Shapiro, Carl, "Patent Licensing and *R&D* Rivalry," *American Economic Review Proceedings*, May 1985, *75*, 25–30.

Spence, A. Michael, "Cost Reduction, Competition and Industry Performance," *Econometrica*, January 1984, *52*, 101–21.

Wilson, Robert, "International Licensing of Technology: Empirical Evidence," *Research Policy*, April 1977, *6*, 114–26.

U.S. Department of Commerce, *Survey of Current Business*, Washington: USGPO, 1985.

Two-Moment Decision Models and Expected Utility Maximization

By JACK MEYER*

Two-moment decision models are consistent with expected utility maximization only if the choice set or the agent's preferences are restricted. This paper identifies a restriction which is sufficient to ensure this consistency and confirms that it holds in many economic models. The implications for economic analysis are then derived.

Two different approaches to representing an agent's preferences over strategies yielding random payoffs are in wide use. Under the mean-standard deviation (MS) approach, the agent is assumed to rank the alternatives according to the value of some function defined over the first two moments of the random payoff, while the expected utility (EU) criterion assumes that the expected value of some utility function defined over payoffs is used instead. The fact that there are these two competing approaches has generated a considerable literature. Some authors are concerned with the advantages and disadvantages of each, while others deal with conditions under which the potentially different approaches would yield the same results, or at least approximately so. Space does not permit this vast literature to be reviewed here, but the following anomaly in this literature is the impetus for this paper.

It is well known that some restriction must be placed on either the agents' preferences or the set of random variables comprising the choice set if consistency between an EU and MS ranking of those alternatives is to be ensured. All such restrictions presented in the literature, such as requiring that the agent's utility function be quadratic or that the random alternatives be normally distributed, have serious theoretical defects and/or have no empirical support. This is true in virtually every economic model one examines, and hence these restrictions can be imposed or verified under only the rarest of circumstances. Simultaneous with this lack of reason to predict that an EU and MS analysis of an economic model will yield the same or similar results is a growing body of literature which in fact points out such similarities. This has occurred in purely theoretical studies and in empirical analyses.[1] From this anomaly one can conclude that some condition is likely to exist which is both sufficient to ensure consistency between the EU and MS approaches, and is theoretically supportable and empirically verifiable in at least some common economic models. The main thrust of this paper is the presentation of such a condition and the determination of its implications for analysis of decision models involving randomness. The results obtained can be viewed as ones which improve moment based decision models in at least two ways. First, consistency with EU is ensured under more acceptable conditions, and second, the various hypotheses and assumptions that make EU analysis so powerful are translated into equivalent conditions in the MS framework.

The paper is organized as follows. Section I briefly presents the consistency condition used throughout the paper and notes theoretical support for it. Also, the literature

*Department of Economics, Michigan State University, East Lansing, MI 48824. I thank Daniel S. Hamermesh, Josef Hadar, Lindon J. Robison, and Peter Schmidt for helpful comments and discussion.

[1]See Gabriel Hawawani (1978) and Haim Levy and Harry Markowitz (1979) for two such papers.

directly related to this condition is mentioned. Section II goes on to derive the implications of the condition for preferences in the MS model. Various definitions and hypotheses from the EU approach are translated into equivalent properties of the MS ranking function. In Section III the implications for comparative static analysis are discussed and several models which could be analyzed using MS methods are presented as examples and as evidence supporting the consistency condition.

I. The Location and Scale Parameter Condition

Two conditions are often cited as being sufficient for a ranking of a set of random variables by the EU and MS criteria to be consistent with one another. Each of these however, has well-known theoretical weaknesses and/or is rarely met when tested for. Thus, neither quadratic utility nor normality are considered to be conditions which are acceptable ones to impose. In this section another condition is presented which also implies that an agent's EU ranking of a set of random variables could instead be represented by a ranking based only on their means and standard deviations. This condition, however, is one which has support and is in fact already met without further assumption in a wide variety of economic models. Surprisingly, the condition presented here has been stated before, has been used before, and yet seems to have been misunderstood and ignored. Certainly, it has not been used to full advantage in the analysis of economic models involving randomness.

The location and scale parameter condition (LS) described here uses the following familiar definition from William Feller (1966):

DEFINITION: *Two cumulative distribution functions* $G_1(\cdot)$ *and* $G_2(\cdot)$ *are said to differ only by location parameters* α *and* β *if* $G_1(x) = G_2(\alpha+\beta x)$ *with* $\beta > 0$.

More common terminology is to use the words location *and scale* parameters to describe α and β and that convention is followed here. One can easily present similar definitions which apply to either density or probability distribution functions. Also, random variables whose cumulative distribution functions (CDF) satisfy this definition are said to differ from one another only by location and scale parameters. Simply stated, the consistency condition of interest here is that the choice set be composed of random variables which differ from one another only by location and scale parameters. A more precise formulation of this LS condition is presented in the next section; first, though, a few comments are in order.

Many two-parameter families of random variables that are described in statistics textbooks are made up of members which differ from one another only by location and scale parameters. The normal and uniform families are examples. Other two-parameter families such as the lognormal family do not satisfy this condition. Also, as Michael Rothschild and Joseph Stiglitz (1970) point out, any set of CDFs whose members differ only by location and scale parameters can be said to form a two-parameter family. Thus the LS condition is a well-defined special case of the ill-defined two-parameter family condition. (See Rothschild-Stiglitz for further discussion of this issue.)

Note also that the LS condition is one which applies to the choice set as a whole, specifying how the random variables must be related to one another, but places no restriction on the functional form of the CDF describing any particular random variable. A particular random payoff can be distributed in any fashion whatsoever, as long as all alternatives to which it is to be compared are similarly distributed in the LS sense.

The LS condition was stated by Rothschild-Stiglitz, but not used by them or discussed except in passing, while previous to their work, James Tobin (1958) used the LS condition as it applies to normally distributed random variables in order to demonstrate certain things concerning that family. That the LS condition has been misunderstood or ignored is most obvious when one realizes that a wide variety of economic models have been analyzed using EU meth-

ods when the MS criterion was equivalent, or have assumed normality or quadratic utility when no such assumption was required, or have been analyzed using both the EU and MS criteria without recognizing that the results obtained were necessarily identical. Prominent among these models are Agnar Sandmo's (1971) model of the competitive firm under price uncertainty and its many extensions, Tobin's (1958) model of pure liquidity preference, and Gershon Feder's (1977) "general economic decision model." These economic models and many others display the property that a single-outcome variable depends on choice variables and parameters, one of which is random, and depends linearly on this random parameter. This property is sufficient to ensure that all outcome or payoff variables differ from one another only by location and scale parameters. Further discussion of these models is deferred to Section III.

To summarize, the LS condition is one which is met in a wide variety of economic models due to the structure of the model itself, and is sufficient to ensure that an EU ranking of the elements of the choice set could instead be represented by an MS ordering. The remaining sections of this paper are devoted to a study of the implications of this for analysis of these models involving randomness.

II. Implications for Preference Representative

When an assumption is used to guarantee that MS and EU rankings of a set of random variables are consistent with one another, then that assumption can also be used to develop relationships between the preference representations in the two approaches. For example, Tobin (1958), and later David Baron (1977), note that assuming a quadratic utility function implies that the MS ranking function $V(\sigma,\mu)$ is of the form $\mu+b(\sigma^2+\mu^2)$ and hence yields concentric semicircles as indifference curves in (σ,μ) space.

Similarly John Chipman (1973) shows that assuming that the random payoffs are all normally distributed implies that $V(\sigma,\mu)$ satisfies the differential equation $(1/\sigma)(\partial V/\partial\sigma)=(\partial^2 V/\partial\mu^2)$. In each of these instances the strong restrictions on $V(\sigma,\mu)$ which result from the consistency assumption are viewed as arguments against the assumption. In this section, the implications of the LS condition for $V(\sigma,\mu)$ are examined in considerable detail. It is pointed out that the resulting $V(\sigma,\mu)$ function is quite flexible, and that many of the restrictions the LS condition places on it are ones that are desirable, and have been assumed in the past without necessarily having a basis for such assumptions. Furthermore, under the LS condition, various popular and interesting hypotheses concerning absolute and relative risk-aversion measures in the EU setting can be translated into equivalent properties concerning $V(\sigma,\mu)$. This is useful if the two approaches are to be combined in analyzing a particular economic model, or if graphical MS techniques are to be used to illustrate or further explain various standard findings in EU-based economic analysis. Thus, contrary to previous consistency conditions, the LS condition puts a useful rather than overly restrictive structure on preferences in (σ,μ) space.

In order to more formally specify the LS condition, and at the same time establish the relationship it implies concerning the preference representations under the MS and EU approaches, consider the following. Assume a choice set in which all random variables Y_i differ from one another only by location and scale parameters.[2] Let X be the random variable obtained from one of the Y_i using the normalizing transformation $X=(Y_i-\mu_i)/\sigma_i$ where μ_i and σ_i are the mean and standard deviation of Y_i. All Y_i, no matter which was selected to define X, are equal *in distribution* to $\mu_i+\sigma_i X$. Hence, the expected utility from Y_i for any agent with utility function $u(\,)$ can be written as

$$(1)\qquad \mathrm{EU}(Y_i)=\int_a^b u(\mu_i+\sigma_i x)\,dF(x)$$

$$\equiv V(\sigma_i,\mu_i),$$

[2] The means and variances of these random variables are assumed to be finite.

where a and b define the interval containing the support of the normalized random variable X. These can be finite or infinite. It is assumed that this integral converges, which places some boundedness restrictions on $u(x)$. See Kenneth Arrow (1974) and William Russell and Tae Kun Seo (1978) for a discussion of these issues.

As indicated, expression (1) defines the MS preference function associated with any utility function in an EU model. This expression serves as the starting point for all analysis concerning the properties of $V(\sigma,\mu)$. Since no restriction has been placed on the form of $u(x)$ or $F(x)$, substantial flexibility remains regarding the form that function $V(\sigma,\mu)$ can take. Certainly more flexibility than if quadratic utility or normally distributed random variables had been assumed. The derivation of a number of properties of $V(\sigma,\mu)$ follow. Certain of these properties are almost too trivial to mention, some are regularly assumed but without formal basis, and others are new. Together these properties form a set which is sufficient for one to conduct significant comparative static analysis in the MS model, often using hypotheses formulated in the EU setting.

The first step in the analysis of $V(\sigma,\mu)$ is to examine its partial derivatives. These are

$$(2)\quad V_\sigma(\sigma,\mu)=\int_a^b u'(\mu+\sigma x)x\,dF(x)$$
$$=\int_a^b\left(u''(\mu+\sigma x)\int_a^x x\,dF(x)\right)dx;$$

$$(3)\quad V_\mu(\sigma,\mu)=\int_a^b u'(\mu+\sigma x)\,dF(x).$$

PROPERTY 1: *$V_\mu(\sigma,\mu)\geq 0$ for all μ and all $\sigma\geq 0$ if and only if $u'(\mu+\sigma x)\geq 0$ for all $\mu+\sigma x$.*

PROPERTY 2: *$V_\sigma(\sigma,\mu)\leq 0$ for all μ and all $\sigma\geq 0$ if and only if $u''(\mu+\sigma x)\leq 0$ for all $\mu+\sigma x$.*

These two properties are quite obvious and display a relationship between preferences in the EU and MS models which is typically assumed. Property 1 implies that movements in the vertical direction in (σ,μ) space are movements to higher indifference curves. The proof of this property is obvious. Property 2 indicates that under risk aversion as one moves left to right in (σ,μ) space, lower indifference curves are encountered. The proof of this property is also quite simple once one notes that X has a mean of zero which implies that $\int_a^x x\,dF(x)\leq 0$. Rothschild and Stiglitz give an example of a two-parameter family of distribution functions for which Property 2 does not hold. Thus, the LS condition, which identifies a proper subset of all two-parameter families, is one which eliminates the possibility of that rather unusual example.

Together these two properties sign the slope of indifference curves in the MS model. Let $S(\sigma,\mu)$ represent the slope of an indifference curve as it passes through (σ,μ). $S(\sigma,\mu)$ is given by

$$(4)\quad S(\sigma,\mu)=\frac{-V_\sigma(\sigma,\mu)}{V_{\mu(\sigma,\mu)}}$$
$$=\frac{-\int_a^b u'(\mu+\sigma x)x\,dF(x)}{\int_a^b u'(\mu+\sigma x)\,dF(x)};$$

$$(5)\quad =\frac{\int_a^b\left(u''(\mu+\sigma x)\int_a^x x\,dF(x)\right)dx}{\int_a^b u'(\mu+\sigma x)\,dF(x)}.$$

PROPERTY 3: *$S(\sigma,\mu)\geq 0$ for all μ and all $\sigma\geq 0$ if $u'(\mu+\sigma x)\geq 0$ and $u''(\mu+\sigma x)\leq 0$ for all $\mu+\sigma x$.*

Moving on to less obvious properties, the concavity/quasi concavity of $V(\sigma,\mu)$, is considered next. Since $V(\sigma,\mu)$ is the representation of ordinal preferences, it is obvious that its quasi concavity is the relevant property to discuss in order to establish convexity of preferences in (σ,μ) space. On the other hand, there are those who have addressed the question of which $V(\sigma,\mu)$ functions can arise as a result of the calculation in equation (1). That is, which $V(\sigma,\mu)$ arise directly from an EU calculation. For this latter ques-

tion, the fact that $V(\sigma, \mu)$ is concave under appropriate conditions is an important one to be cognizant of. For this reason, and because concavity implies quasi concavity, concavity is the property which is discussed.[3] To do this the following second derivatives of $V(\sigma, \mu)$ are used:

$$(6)\quad V_{\mu\mu}(\sigma,\mu)=\int_a^b u''(\mu+\sigma x)\,dF(x);$$

$$(7)\quad V_{\mu\sigma}(\sigma,\mu)=\int_a^b u''(\mu+\sigma x)x\,dF(x);$$

$$(8)\quad V_{\sigma\sigma}(\sigma,\mu)=\int_a^b u''(\mu+\sigma x)x^2\,dF(x).$$

PROPERTY 4: *$V(\sigma, \mu)$ is a concave function for all μ and all $\sigma \geq 0$ if and only if $u''(\mu+\sigma x) \leq 0$ for all $\mu+\sigma x$.*

One can establish the concavity of $V(\sigma, \mu)$ by showing that $V_{\mu\mu}(\sigma,\mu)$ and $V_{\sigma\sigma}(\sigma,\mu)$ are nonpositive and that $V_{\mu\mu}V_{\sigma\sigma}-V_{\mu\sigma}^2$ is nonnegative. The first two of these sign restrictions follow in an obvious manner from the condition imposed on $u''(\mu+\sigma x)$, the third is slightly more difficult to show and that proof is found in the Appendix.

This concavity and quasi concavity of $V(\sigma, \mu)$ implies that preferences in the MS decision model display the convexity property; that is, all points preferred or indifferent to a given point form a convex set. This property is especially useful in economic models in which $V(\sigma, \mu)$ is to be maximized over feasible sets which are also convex. Tobin (1958) demonstrated this property for normally distributed random variables, and in fact, his proof actually is general enough to establish Property 4. Hence this property is not completely new. Martin Feldstein (1969) points out the fact that not all MS preferences consistent with expected utility maximization are this well-behaved. He constructs an example where all random alternatives are lognormally distributed and $u(x)=\ln x$, in which convexity of preferences fails to hold.[4] Of course, as mentioned earlier, the family of lognormally distributed random variables does not satisfy the LS condition. Thus, Feldstein's example is not a counterexample to this result, but it does point out one of the restrictions placed on $V(\sigma, \mu)$ by the condition used here. In this case, the restriction may well be a desirable one. This convexity property was the prime consideration in choosing to formulate this research in terms of mean and standard deviation rather than mean and variance. Mean-variance decision models, even for the case of normally distributed random alternatives, do not display convexity of preferences for broad classes of utility functions.

The next property to be discussed is one which has been assumed by others, sometimes supported with examples in which it holds, but has not been demonstrated as following from a general set of assumptions. It is a property which is extremely valuable in conducting comparative static analysis of economic problems and worthy of formal demonstration.

PROPERTY 5: *$\partial S(\sigma,\mu)/\partial\mu \leq (=, \geq) 0$ for μ and all $\sigma > 0$ if and only if $u(\mu+\sigma x)$ displays decreasing (constant, increasing) absolute risk aversion for all $\mu+\sigma x$.*

The proof of this property is in the Appendix. The property indicates how the slope of an indifference curve changes as one moves in the vertical (μ) direction. Notice that no matter how the random variable is distributed, the slope of the agent's risk-aversion measure is the determining factor. An assumption of constant absolute risk aversion on the part of the agent in the EU model implies that indifference curves in the MS model are vertically parallel.

[3]As a referee has pointed out, quasi concavity is easily established directly.

[4]Feldstein (p. 8) also mistakenly claims that Property 4 does not hold for the pure liquidity preference model. His constructed example is faulty in that a corner solution is obtained due to a borrowing restriction rather than nonconvexity of preferences. "Plungers" are not possible in this simple portfolio model.

This property of $V(\sigma, \mu)$ was assumed by both Michael Adler (1969) and Gabriel Hawawani (1978) in their MS analysis of the liquidity preference and competitive firm models. Adler provides three examples in which the property holds, but does not prove it holds in general. Hawawani argues that the slope of an indifference curve in (σ, μ) space is a measure of risk sensitivity and hence decreasing, constant or increasing risk aversion should be *defined* by the sign of $\partial S(\sigma, \mu)/\partial \mu$. As each of these researchers show, this property is extremely useful in drawing conclusions concerning the effects of various shifts in the opportunity set in (σ, μ) space. This is particularly true for vertically parallel shifts in the opportunity set, shifts which occur in the firm model as fixed costs change, and in the portfolio model as initial wealth is altered.

The above relationship between the absolute risk-aversion measure in the EU model and the slope of indifference curves in the MS model leads one naturally to examine whether other such relationships can also be found. This section concludes with two such relationships whose proofs are in the Appendix.

PROPERTY 6: $\partial S(t\sigma, t\mu)/\partial t \geq (=, \leq) 0$ *if and only if* $u(\mu + \sigma x)$ *displays increasing (constant, decreasing) relative risk-aversion for all* $\mu + \sigma x$.

This property indicates that the slope of indifference curves in (σ, μ) space change as one moves out along a ray in a manner which depends on the relative risk-aversion properties of the agent. If one accepts the hypothesis that agents display decreasing absolute, but increasing relative risk aversion, then the implication in these MS models is that indifference maps are somewhere between those which are vertically parallel and those where the slopes are constant along rays.

The final property to be formally presented considers the implication of the concept of more risk averse in the Pratt-Arrow sense on preferences in (σ, μ) space.[5] As one would hope, the EU-based definition translates into steeper sloped indifference curves in the MS model. The formal statement of this follows and the proof is in the Appendix.

PROPERTY 7: $S_1(\sigma, \mu) \geq S_2(\sigma, \mu)$ *for all* (σ, μ) *if and only if* $u_1(\mu + \sigma x)$ *is more risk averse than* $u_2(\mu + \sigma x)$ *for all* $\mu + \sigma x$.

It is obvious that one can use the framework developed here to attempt to find other properties which might prove useful in MS modeling. For instance, it is easy to show that $S(0, \mu) = 0$, that is, indifference curves in the MS model are flat as they touch the vertical axis. The properties demonstrated so far, however, appear to be the most important ones, and the ones which have received the most attention in the literature.

III. Comparative Statics

One of the uses of the results presented in previous sections is to justify using simple two-dimensional MS methods to conduct a comparative static analysis of an economic model in which the LS condition holds. This might be done to augment, illustrate, or replace an EU-based analysis of the same model. While simplicity and expository reasons are the main rationale for doing this, the use of an alternate methodology can also allow the investigator to recognize results which might not be obvious in more complicated frameworks. Since comparative static analysis under the MS approach is straightforward and uses standard calculus techniques, only a brief sketch of the procedure involved is presented here. A listing of several models in which the procedure can easily be applied is also given.

Many if not most economic models involving randomness are structured so that there is a single outcome variable whose value is written as a function of the choices made by the agent and various parameters describing the economic environment. One of these parameters is assumed to be random or unknown and hence is represented by a random variable. Whenever the model specifies the outcome variable as a positive linear function of this random parameter,

[5]The property was suggested by a referee and by Josef Hadar.

then all possible outcome variables resulting from the choices the agent can make differ from one another only by location and scale parameters.[6] It is this category of models that is considered in this section. In addition to being able to use MS methods to represent the agent's choices in these models, one can also examine the effects of changing any nonrandom parameter or changing the random parameter to one which differs only by location and scale parameters. This is because these changes all generate outcome alternatives which differ from those originally available, and each other, only by location and scale parameters. The effect of a completely arbitrary change in the random parameter cannot be analyzed.[7]

Although there are several ways to formulate a general economic model of the type described above, the one used here is quite natural to economists due to its similarity to the standard, two-good, consumer model. It is assumed that the agent maximizes $V(\sigma, \mu)$ subject to restrictions defining the set of (σ, μ) from which he or she can choose. These restrictions can always be represented parametrically by calculating the mean and standard deviation of the outcome variable as functions of the parameters d and choice variables α.

Thus the problem is to choose α to maximize $V(\sigma, \mu)$ where $\sigma = \sigma(d, \alpha)$ and $\mu = \mu(d, \alpha)$. It is also often the case that one can eliminate one parameter or choice variable and arrive at a single constraint of the form $\mu = m(d, \alpha, \sigma)$. Solving either of these constrained maximization problems is a straightforward exercise.

This section is concluded by briefly describing several important economic models in which the just described comparative static analysis can be carried out. Sandmo introduced a model of the competitive firm in which output level is selected by the firm taking random output price as given and described by $F(p)$. Output is selected so as to maximize expected utility from profits which are given by $\pi = px - c(x) - B$. Variable and fixed costs are $c(x)$ and B, respectively. This economic model is formulated so that all random profit alternatives available to the firm are positive linear transformations of the given random variable p, and hence are related to one another by location and scale parameters. Thus, the expected utility-maximizing choice of x can instead be represented as one which maximizes $V(\sigma, \mu)$ subject to an opportunity set described by

$$\mu = \mu_p x - c(x) - B; \tag{9}$$

$$\sigma = \sigma_p x. \tag{10}$$

Solving out for x in the second equation and substituting in the first, one obtains an expression giving μ as a function of σ and the parameters of the problem.

$$\mu = (\mu_p/\sigma_p)\sigma - c(\sigma/\sigma_p) - B. \tag{11}$$

Most of the extensions of Sandmo's work also display this simplifying characteristic.

Similarly in Tobin's pure theory of liquidity preference, which is also the portfolio model involving a single risky and riskless asset, terminal wealth is given by $W = \alpha W_0 \rho + (1-\alpha) W_0 r$ where α is the proportion of initial wealth W_0 invested in the riskless asset, and ρ and r are the returns to the riskless and risky assets. The return to the risky asset is assumed to be random and described by CDF $F(r)$. All final wealth random variables are related to one another and random parameter r only by location and scale parameters. Hence, no matter how r is distributed and no matter which utility function is specified in the EU model, the expected utility-maximizing choice of α can be represented as one which maximizes $V(\sigma, \mu)$ subject to the constraints in the problem. The opportunity set in (σ, μ) space is given by

$$\mu = \alpha W_0 \rho + (1-\alpha) W_0 \mu_r; \tag{12}$$

$$\sigma = (1-\alpha) W_0 \sigma_r, \tag{13}$$

[6] One can also use a negative linear transformation, but cannot use one which is sometimes positive and sometimes negative.

[7] It is interesting to note that many comparative static results in EU models are also for location and scale transformations of the random variable. Prominent among these is the risk-increasing transformation $x + \theta(x - \bar{x})$.

where μ_r and σ_r are the mean and standard deviation of r. This is easily rewritten as the linear restriction

$$\mu = W_0\rho + ((\mu_r - \rho)/\sigma_r)\sigma \tag{14}$$

where $\sigma \in [0, \sigma_r]$.

Finally, a general decision model, developed to include in it many economic models frequently encountered by economists, was presented by Feder. This model assumes that an agent maximizes expected utility from $\theta\psi(x) + \phi(x) + A$ where θ is a random variable, x is a vector of choice variables, $\psi(x)$ and $\phi(x)$ are real valued functions and A is a constant. An equivalent formulation of this general decision model is that vector x is selected to maximize $V(\sigma, u)$ where $\mu = \mu_\theta\psi(x) + \phi(x) + A$ and $\sigma = \psi(x)\sigma_\theta$ are the restrictions defining the choice set. For analysis of these and many other models, the results of this paper are relevant.

IV. Summary and Conclusions

This paper has reidentified and emphasized the location and scale parameter condition as one which is sufficient to ensure consistency between expected utility and moment-based rankings of random variables. Moreover, this condition is shown to be one which does hold in many economic models, and thus provides the explanation for unpredicted identical findings noted by researchers using the two supposedly different analysis methodologies.

A number of issues remain for further research. First, in order to explain the similarities in findings in empirical studies using EU and MS techniques, one must be able to empirically test for the LS condition. Preliminary work indicates that this is a rather difficult statistical problem whose resolution is likely to involve Kolmogorov-Smirnov type statistics. It is important to note however that since perfectly correlated random variables satisfy the LS condition, well-diversified portfolios are likely to satisfy statistical tests for the condition, and this provides a potential explanation for findings such as those of Haim Levy and Harry Markowitz (1979).

Another area requiring further work involves use of nonlinear two-parameter transformations of the random parameter to generate choice sets which can be ranked consistently using MS or EU methods. Such transformations as $e^{\alpha+\beta X}$ certainly yield the required consistency, but more work is needed to establish their usefulness. This exponential transformation, when applied to a normally distributed parameter, yields lognormally distributed outcome variables. Its convex nature explains Feldstein's "plunging" indifference curves finding.

APPENDIX

Necessity of the condition on $u(\mu+\sigma x)$ in Properties 1, 2, 4, 5, 6, and 7 is demonstrated indirectly. In each case, one assumes the condition on $u(\mu+\sigma x)$ is violated in some interval, and then for X with bounded support chooses parameters μ and σ so that the support of random variables $\mu + \sigma X$ lie entirely within the offending interval. For these μ and σ, the stated condition on $V(\sigma, \mu)$ would not hold. Formal presentation of these necessity proofs is not contained in the demonstrations which follow.

THEOREM 1: *$V_{\mu\mu}V_{\sigma\sigma} - V_{\mu\sigma}^2 \geq 0$ for all μ and all $\sigma \geq 0$ if and only if $u''(\mu+\sigma x) \leq 0$ for all $\mu + \sigma x$.*

PROOF: To establish this let x^* be defined by

$$x^* \int_a^b u''(\mu+\sigma x)\,dF(x) = \int_a^b u''(\mu+\sigma x)\,x\,dF(x),$$

thus $\int_a^b (x - x^*)u''(\mu+\sigma x)\,dF(x) = 0$. The integrand of this expression changes sign once from positive to negative. This implies that $\int_a^b x(x-x^*)u''(\mu+\sigma x)\,dF(x) \leq 0$. Rewriting this as $\int_a^b u''(\mu+\sigma x)x^2\,dF(x) \leq x^*\int_a^b u''(\mu+\sigma x)x\,dF(x)$ and multiplying both sides by $\int_a^b u''(\mu+\sigma x)\,dF(x)$ yields the desired result.

PROOF OF PROPERTY 5:

$$\frac{\partial S}{\partial \mu}=\frac{-\int u'dF\int u''x\,dF+\int u'x\,dF\int u''\,dF}{\left(\int u'dF\right)^2}$$

where the argument of u' and u'', $\mu+\sigma x$ has been suppressed for notational convenience. The sign of $\partial S(\sigma,\mu)/\partial\mu$ is the same as the sign of the numerator of this expression. Let x^* satisfy $x^*\int u'dF=\int u'x\,dF$. Thus $\int(x-x^*)u'dF=0$ and the integrand changes sign once from negative to positive. Furthermore, the expression to be signed can be written as

$$\int u'dF\int ru'x\,dF-\int u'x\,dF\int ru'\,dF,$$

where $r(\mu+\sigma x)=-u''(\mu+\sigma x)/u'(\mu+\sigma x)$. Using the definition of x^* this reduces to

$$\left[\int u'dF\right]\left[\int r(x-x^*)u'dF\right].$$

Since $\int u'dF\geq 0$, this expression is nonpositive, zero or nonnegative as $r(\mu+\sigma x)$ is a decreasing, constant or increasing function of x.

PROOF OF PROPERTY 6:

Since (σ,μ) is arbitrary signing $\partial S/\partial t|_{t=1}$ is sufficient.

$$\frac{\partial S}{\partial t}\Big|_{t=1}=\left[-\int_a^b u'dF\int_a^b u''x(\mu+\sigma x)\,dF\right.$$
$$\left.+\int_a^b u'x\,dF\int_a^b u''(\mu+\sigma x)\,dF\right]$$
$$\Big/\left(\int_a^b u'dF\right)^2.$$

Now the relative risk-aversion measure at $\mu+\sigma x$ is

$$K(\mu+\sigma x)=\frac{-u''(\mu+\sigma x)}{u'(\mu+\sigma x)}(\mu+\sigma x).$$

Thus

$$\frac{\partial S}{\partial t}=\left[+\int_a^b u'\,dF\int_a^b Ku'x\,dF\right.$$
$$\left.-\int_a^b u'x\,dF\int_a^b Ku'\,dF\right]\Big/\left(\int_a^b u'\,dF\right)^2,$$

but this expression is identical to that found in the proof of Property 5 and hence the theorem follows in the same manner.

PROOF OF PROPERTY 7:

Assume $u_1(\mu+\sigma x)=v(u_2(\mu+\sigma x))$ where v is concave and increasing. This implies that

$$S_1(\sigma,\mu)=-\int_a^b v'u_2'x\,dF\Big/\int_a^b v'u_2'dF$$

Thus one must show that

$$\frac{-\int_a^b v'u_2'x\,dF}{\int_a^b v'u_2'dF}\geq\frac{-\int_a^b u_2'x\,dF}{\int_a^b u_2'\,dF}$$

or

$$\int_a^b v'u_2'x\,dF\int_a^b u_2'\,dF\leq\int_a^b v'u_2'\,dF\int_a^b u_2'x\,dF.$$

The concavity of v and the method used in proving Property 5 can now be used to complete the proof.

REFERENCES

Adler, Michael, "On the Risk-Return Trade-off in the Valuation of Assets," *Journal of Financial and Quantitative Analysis*, December 1969, *4*, 493–512.

Arrow, Kenneth J., "Use of Unbounded Utility Functions in Expected Utility Maximization-Response," *Quarterly Journal of Economics*, February 1974, *88*, 136–38.

Baron, David P., "On the Utility Theoretic Foundations of Mean-Variance Analysis," *Journal of Finance*, December 1977, *32*, 1683–97.

Chipman, John S., "The Ordering of Portfolios in Terms of Mean and Variance," *Review*

of Economic Studies, April 1973, *40*, 167–90.

Feder, Gershon, "The Impact of Uncertainty in a Class of Objective Functions," *Journal of Economic Theory*, December 1977, *16*, 504–12.

Feldstein, Martin S., "Mean-Variance Analysis in the Theory of Liquidity Preference and Portfolio Selection," *Review of Economic Studies*, January 1969, *36*, 5–12.

Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. II, New York, Wiley & Sons, 1966.

Hawawani, Gabriel A., "A Mean-Standard Deviation Exposition of the Theory of the Firm under Uncertainty: A Pedagogical Note," *American Economic Review*, March 1978, *68*, 194–202.

Levy, Haim, "The Rationale of the Mean-Standard Deviation Analysis: Comment," *American Economic Review*, June 1974, *64*, 434–41.

______ **and Markowitz, H.,** "Approximating Expected Utility by a Function of Mean and Variance," *American Economic Review*, June 1979, *69*, 308–17.

Rothschild, M. and Stiglitz, J. E., "Increasing Risk: 1. A Definition," *Journal of Economic Theory*, September 1970, *2*, 225–43.

Russell, W. R. and Seo, T. K., "Admissible Sets of Utility Functions in Expected Utility Maximization," *Econometrica*, January 1978, *46*, 181–84.

Sandmo, Agnar, "On The Theory of the Competitive Firm under Price Uncertainty," *American Economic Review*, March 1971, *61*, 65–73.

Tobin, James, "Liquidity Preference as Behavior Towards Risk," *Review of Economic Studies*, February 1958, *25*, 65–86.

Trade in a Tiebout Economy

By JOHN DOUGLAS WILSON*

This paper examines interregional commodity trade in a "Tiebout Economy," that is, a many-region economy with perfect labor mobility and endogenous government decision making. For a model with scale economies in public good consumption, it is shown that any equilibrium is both Pareto efficient and asymmetric: regions containing the same types of individuals and production possibilities nevertheless differ in the traded goods which they produce and the public good levels which they provide residents.

The purpose of this paper is to explore both the positive and normative aspects of commodity trade in a Tiebout economy. By "Tiebout economy," I mean an economy with a large number of small "regions" (for example, communities or states) and no impediments to the mobility of individuals between these regions. This type of economy has received widespread attention since the appearance of Charles Tiebout's classic 1956 article on local public goods, in part because it bears a close relation to the standard competitive model. This relation suggests that Tiebout economies may serve as an important "benchmark case," around which the behavior of actual systems of local governments can be better understood. Yet there has been almost no analysis of the role of interregional trade in this type of economy. Eitan Berglas (1976) provides an important exception, which I later discuss.

Here I present a model in which trade emerges as a general equilibrium response of the economy to scale economies in public good consumption. In fact, each local government reacts to these scale economies by choosing a tax and expenditure policy under which the private production sector produces only a single good which is traded between regions. As a result of this specialization, the model has the surprising property that regions containing the same types of individuals and production possibilities nevertheless differ in both the traded goods which are produced and the output levels of public goods. This conclusion extends Joseph Stiglitz' (1977) finding that, for a system of nontrading regions, the nonconvexities associated with public goods make possible the existence of an equilibrium where wages and public good levels differ across identical regions, although all individuals are identical. The present paper shows that such differences *must* occur when regions trade. Furthermore, these differences necessarily occur even though the number of regions is arbitrarily large, in which case the scale economies associated with public good provision in any one community are small relative to the economy as a whole.

Despite the demonstration that consumption bundles differ across identical individuals in equilibrium, I shall show that an equilibrium is Pareto efficient. Differences in consumption bundles are shown to play a vital role in the achievement of an efficient allocation.

The plan of the paper is as follows. In Section I, a simple model with "pure" public goods is presented. Section II examines the role of commodity trade in this model, and Section III proves that the equilibrium is Pareto efficient. Congestion is added to the model in Section IV, and scale economies

*Department of Economics, Indiana University, Bloomington, IN 47405. This is a revised version of a paper I presented at the Sloan Workshop in Urban Public Economics at the University of Maryland, the McMaster-Queen's Workshop on Systems of Cities, and the 1986 North American Summer Meeting of the Econometric Society. I am grateful to the conference participants, the referees, Richard Arnott, Eitan Berglas, David Wildasin, and especially David Pines for helpful comments and suggestions.

are again shown to produce commodity trade between regions. Several other extensions, and some limitations, are also described in Section IV. The final section draws some comparisons between the results reported here and the international trade literature.

I. The Model

To highlight the role of public goods as a basis for trade, I consider a model where there are no innate differences between individuals or regions. The economy contains a large number of identical "regions," each possessing the same fixed amount of land. Each individual occupies a single region, where he supplies one unit of labor and consumes goods. For simplicity, I assume that there are only two private goods and a single public good. The private goods are produced by competitive private firms using land and labor, and the public good is produced using the private goods as inputs. The technologies for both private and public good production exhibit constant returns to scale and are identical across regions.[1] The private goods are both tradeable between regions, and each region is small in the sense that it has a negligible impact on the terms of trade. For notational ease only, the public good is assumed to be nontraded, meaning it is consumed in the region where it is produced.

The public good is a Samuelson pure public good in the sense that there is no rivalry in consumption: once G units are available to one resident of a region, the same G units are available to all residents. Thus, additional consumption economies can be achieved in public good provision by adding additional residents to the region. (It is the production of the G units which exhibits constant returns to scale.) However, adding these additional residents to the region's fixed land supply lowers the marginal product of labor. This tradeoff between consumption economies and diminishing labor productivity determines the equilibrium size of regions. Much more general assumptions about the public good will be introduced in Section IV.

The economy contains three types of "agents": consumer-workers ("individuals"), private good producers, and land developers. I next discuss the maximization problem facing each group.

An individual's utility depends on his consumption of the private goods and public good, and he treats the public good supply as fixed when choosing his utility-maximizing private good demands. The demand for each good is assumed to be positive at all prices (i.e., indifference surfaces are strictly convex and converge asymptotically to each of the three axes). All individuals possess identical utility functions and supply one unit of labor.[2] They may possess different amounts of land in different regions, but such differences are irrelevant for two reasons. First, there is perfect labor mobility between regions, implying that each individual's residence is completely independent of where he owns land. Second, regions are "developed" (i.e., occupied by residents and firms) until the net return to land in each region is driven to zero. To obtain this property, I assume that the number of regions is so large that it never becomes a binding constraint on potential development. Thus, unoccupied regions usually exist in equilibrium.[3]

Each individual possesses an indirect utility function, $u(Y, G, p)$, where Y is private after-tax income, G is the public good supply, and p is the vector of private good prices, $p = (p^1, p^2)$. With Y_i and G_i denoting income and the public good supply available to an individual in region i, the locational problem facing the individual may be ex-

[1] The production functions are assumed to be strictly quasi concave and continuously differentiable. Utility functions are also assumed to possess these properties.

[2] An individual's labor supply could be included in the utility function as a choice variable without affecting the results.

[3] My working paper (1986) shows that, as long as developers behave "competitively" in the manner specified below, the results are not altered in cases where a binding constraint on the number of available regions creates positive net land rents in equilibrium. Suzanne Scotchmer (1986, Appendix C) examines a model of this type without interregional commodity trade.

pressed algebraically as follows:

$$(1) \qquad \operatorname*{Max}_{i} u(Y_i, G_i, p).$$

In equilibrium, all individuals receive the same utility level.

The problem facing a good i producer is to choose labor and land inputs, L_i and T_i, to maximize profits. In equilibrium, maximum profits equal zero:

$$(2) \qquad \operatorname*{Max}_{L_i, T_i} p_i f_i(L_i, T_i) - wL_i - rT_i = 0,$$

where f_i is a production function exhibiting constant returns to scale, w is the wage, and r is the rental rate on land. I shall assume that industry 2 is always labor intensive relative to 1; meaning, $L_2/T_2 > L_1/T_1$.

I assume (and justify shortly) that each region's tax and public expenditure policy is controlled by a land developer (or land management company) who represents the interest of the region's landowners. The developer's objective is to maximize the net return to land.[4] To determine this net return, I first let p_G denote the minimum unit cost of the public good. This cost is determined by the price vector, p, independently of G, since the public good has been assumed to be produced from the two private goods via a constant returns to scale technology. Given p_G, the total tax which a region's landowners must pay to finance the public good is $p_G G - bL$, where b is a head tax on residents and L is the region's labor supply. By subtracting this tax from the total gross return to land, I obtain the (total) net return: $R = rT + bL - p_G G$, where T is the region's land supply. In equilibrium, the maximum R equals zero.

It is useful to initially let b, G, and L serve as the control variables for the developer's maximization problem. (The graphical argument in the next section shows that the developer need only exercise indirect control over L through his choice of b and G.) Given the region's land supply and the equilibrium product prices, p^*, the region's equilibrium factor prices can be written as functions of $L, r = r(L)$ and $w = w(L)$. The developer then maximizes $R = r(L) \cdot T + bL - p_G G$, subject to a "migration constraint," requiring that each resident obtain at least the equilibrium utility available in other regions, u^*:

$$(3) \qquad u(w(L) - b, G, p^*) \geq u^*.$$

In effect, the assumed "smallness" of regions implies that the developer treats labor as being infinitely elastic at the equilibrium utility level u^*.

The prominent role given to developers here is at least partially justified by the observation that a region's residents agree unanimously with the developer's choices when the economy is in equilibrium. Stiglitz (1983) provides a similar unanimity result for an economy with different types of workers, but no commodity trade. To understand the result in the present context, suppose to the contrary that, with the economy in equilibrium, there exists a feasible policy change for a given region which raises the utilities of residents above u^*. To be feasible, this policy change must provide landowners with a nonnegative net return on land, since they have the option to withold their land from production. If the policy change is implemented, however, there will obviously exist a reduction in G which is small enough not to induce residents to move elsewhere, where they receive only the utility u^*. The cost saving obtained from this reduction would raise the net return to the region's land above zero, thereby contradicting the equilibrium condition that the maximum net return equals zero.

II. Equilibrium Trade

This section demonstrates that each region produces only a single traded good when the economy is in equilibrium. Since the total demand for each good is assumed to be

[4] Models with land value maximization have been extensively studied and found to possess desirable efficiency properties. Some early analyses of land value maximizing communities are found in Elhanan Helpman and David Pines (1977) and Jon Sonstelie and Paul Portney (1978). David Wildasin (1986, sec. 4.3) provides a review of the literature.

always positive, the surprising implication of this result is that some regions produce only good 1 while others produce only good 2, although all regions contain completely identical individuals and production possibilities. As a result of this specialization, public good levels and factor prices differ across regions.

To emphasize the economics behind these results, the arguments in this section will be kept largely verbal and graphical, leaving the more mathematical treatment to Section IV. The essential insight to be developed here is that, when a region produces both goods, there is *no* tradeoff between scale economies in public good consumption and diminishing marginal productivity of labor in private good production. In particular, an increase in a region's population can be obtained without any reduction in the marginal productivity of labor, simply through a transfer of land from good 1 production to good 2 production. For this reason, a developer will never tolerate production of both goods.

I begin by describing two important properties of a developer's optimal policy, both of which are formally proved in Section IV. First, the developer sets his head tax (b) equal to the marginal congestion cost created by an additional resident. Since the present model contains no congestion costs, however, b equals zero.[5] Second, the developer sets his public good output at a level where the total income residents are willing to give up for another unit of G equals the cost of producing that unit. Only then will the developer be maximizing the net return to land, given that residents must receive at least the utility level available elsewhere. Letting MRS_{Gb} denote a resident's marginal rate of substition between public good output and the head tax, I can write this condition symbolically as follows: $L \cdot MRS_{Gb} = p_G$. This is the well-known Samuelson condition for efficient public good provision.

[5]Thus, all public expenditures are financed by land taxes; and, since after-tax land rents equal zero in equilibrium, public good expenditures exactly equal gross land rents in equilibrium. Stiglitz (1977) shows that this is a condition for Pareto efficiency in an economy without trade. (William Vickrey, 1977, presents a related result.) Richard Arnott and Stiglitz (1979) refer to this result as the "Henry George Theorem" and greatly generalize the conditions under which it holds.

It is now a simple manner to show that each region produces only one traded good in equilibrium. Suppose, to the contrary, that a region produces both goods under the developer's optimal policy. Then the region's overall labor-land ratio is a weighted average of the labor-land ratios in each industry, where the weights are both positive represent each industry's share of land:

$$(4) \qquad L/T = (1-\beta)(L_1/T_1) + \beta(L_2/T_2); \quad \beta = T_2/T,$$

where $L_2/T_2 > L_1/T_1$ by assumption. As L rises, factor market equilibrium is maintained by raising β, rather than by changing the factor prices (i.e., the factor price functions defined in the previous section, $r(L)$ and $w(L)$, are constant over the interval of L's where both goods are produced). With $b = 0$, as required for optimality, the rise in L has no impact on the net return to land, $R = rQ + bL - p_G G$. However, the rise in L does raise $L \cdot MRS_{Gb}$ above p_G, since $L \cdot MRS_{Gb} = p_G$ initially, as required for optimality. This means that the total head tax payments residents are willing to make for a marginal increase in G exceeds the cost of this increase: $Ldb > p_G dG$. In other words, it is possible to raise both G and b so that R rises while continuing to provide residents with the equilibrium utility level. (By similar reasoning, the developer could also feasibly raise R by reducing L and then lowering both G and b.) It follows that R cannot be maximized when both goods are produced.

A useful way to graphically illustrate the equilibrium, while providing an alternative proof of the specialization result, is to define the "bid-rent function" for each industry, $r = r_{Bi}(G)$ for industry i. This function gives the maximum rental rate that the industry is willing to pay for land in a region at each level of the public good supply (i.e., the rental rate at which profits equal zero). To construct it, I first use the migration constraint (inequality (3)) to define the wage as a function of $G, w = w(G)$. This function describes the minimum wage at which indi-

viduals are willing to reside in the region at each level of G, given the equilibrium product prices and utility level. With product prices fixed at their equilibrium levels, the zero profit constraint (equality (2)) defines the rental rate as a function of the wage for each industry, $r = r_i(w)$ for industry i. Substituting $w(G)$ into $r_i(w)$ then gives the bid-rent function, $r_{Bi}(G) = r_i(w(G))$, which is illustrated in Figure 1 for each industry.[6] Since $dr_i/dw = -L_i/T_i$, the assumption that industry 2 is relatively labor intensive implies that industry 2 has the more steeply sloped bid-rent function where the two functions cross.

Given a region's chosen G, the equilibrium rental rate on land is the maximium of $r_{B1}(G)$ and $r_{B2}(G)$. Thus, the net return to land is related to G according to the function,

$$(5) \quad R(G) = \left[\max_i r_{Bi}(G)\cdot T\right] - p_G G.$$

The developer's maximization problem consists of choosing G to maximize $R(G)$.

Only at that G where $r_{B1}(G) = r_{B2}(G)$, denoted G^* is Figure 1, will both industries be willing to produce in the region. Elsewhere, the industry with the lower bid-rent function cannot earn nonnegative profits in the region. For G^* to maximize $R(G)$, however, the slope of each bid-rent function must equal p_G at G: Otherwise, there would exist a small change in (r, G) along one of the bid-rent functions which raises the net return to land. Since the bid-rent functions have different slopes at G^*, however, this condition must fail. Hence, I have again shown that each region is completely specialized in equilibrium.

[6]These bid-rent functions need not be concave everywhere. For this reason, there need not always exist a positive and finite G which solves the land value maximization problem. This possibility of nonexistence also arises in models without trade (see Anthony Atkinson and Stiglitz, 1980, Fig. 17–3b). I view the nonexistence of a solution as representing a misspecification of the costs and benefits of increased community size, rather than an interesting economic phenomenon. One assumption which can be shown to insure existence is that the production of each good is "bounded": given the land (labor) employed by a region to produce the good, output converges to a finite upper bound as the labor (land) goes to infinity.

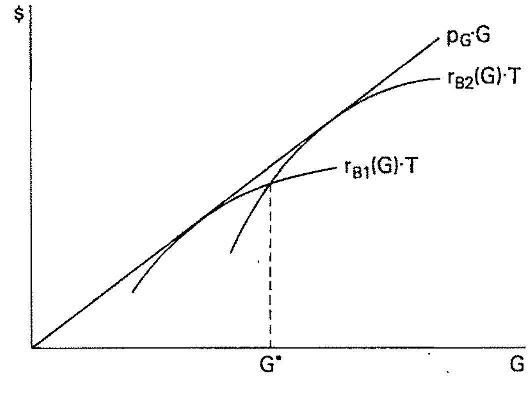


FIGURE 1

Since the maximum R equals zero in equilibrium, the G which a developer chooses must be located where an industry's bid-rent function is tangent to the cost line $p_G G$. As shown, there are two G's with this property. Regions with the low G possess a low r and high w, implying that they attract the land-intensive industry. Regions with the high G have a high r and low w, and attract only the labor-intensive industry.

The double tangency in Figure 1 is achieved through adjustments in the product prices and the utility level, both of which each developer treats as exogenously fixed. Since only relative prices matter, p_1 may be set equal to one, leaving p_2 and u as the two "unknowns." Given u, a sufficiently high (low) p_2 insures that R is maximized only at a G where industry 2 (1) is willing to produce. With p_2 held fixed, a sufficiently high (low) u insures that the maximum R is negative (positive). Using the continuity properties of the model, it can be shown that there exists a unique utility u^* and price p_2^* at which developers achieve a maximum R equal to zero by attractiving either industry. Graphically, this means that the double tangency condition holds. Once it is achieved, supply is equated with demand in each of the economy's two product markets by adjusting the fraction of regions producing each good.[7]

[7]Since the fraction of regions producing each good is not a continuous variable, the equilibrium conditions can only "approximately" hold. But such approxima-

This specialization result would clearly remain valid if regions possessed different technologies or land endowments. The assumption of identical individuals could also be dropped; and, as Section IV shows, the properties of public good production could be greatly generalized. What makes the specialization result particularly interesting, however, is its implication that there must exist interregional commodity trade in the economy, even though *no* innate differences exist between regions or individuals.

III. Pareto Efficiency

The lack of a tradeoff between public good scale economies and diminishing marginal productivity under incomplete specialization also implies that specialization is a necessary characteristic of a Pareto-efficient allocation for this economy. The basic argument may be sketched as follows. If all regions were incompletely specialized under a Pareto-efficient allocation, then the marginal products of labor and land would have to be equated across regions (i.e., there would be factor price equalization). But then a central planner could reallocate labor between any two regions without changing these marginal products. This reallocation would have no impact on the total outputs of either good: production of the labor (land) intensive good would rise in the region receiving more (less) labor and drop by an equal amount in the other region. Thus, it would be technologically feasible to provide each individual with the same consumption bundle which he received before the change. However, the total willingness to pay for the public good, summed over all residents ($L \cdot MRS_{Gb}$), would obviously rise in the region with additional residents and drop in the region with fewer individuals. The planner could then raise every individual's utility by increasing G in the former region and reducing G in the latter region, with the net wages being adjusted to keep utilities equated.

A similar type of argument is made by Berglas. But his model contains two types of labor as the two factors, rather than land and homogeneous labor. Starting from an allocation with factor price equalization and incomplete specialization, Berglas shows that a central planner can raise everyone's utility by creating differences in input ratios across regions and then "tailoring" each region's public good supply to its new population mix. He explains: "The gains from trade among communities arises not from increase in productivity but from increase in consumption efficiency, by making it possible for more homogeneous communities to exist" (p. 420). Thus, trade in his model is a result of demand considerations, rather than the technological considerations considered here.

Not only are both the equilibrium and efficient allocations characterized by specialization, but the equilibrium allocation satisfies all other conditions for efficiency. To prove that an equilibrium is Pareto efficient, I utilize my earlier observation that land value maximization is equivalent to utility maximization. To state the utility-maximization problem in a convenient form, I note that the developer's ability to use both land and head taxes effectively provides him with complete control over the region's labor supply, private good outputs, and private and public consumption levels. With these variables chosen as control variables, the utility maximization problem may be stated as follows:

$$\operatorname{Max} u(C_1, C_2, G)$$

subject to

$$(6)\quad p_1(X_1 - C_1 L - D_1) + p_2(X_2 - C_2 L - D_2) \geq 0$$

$$(7a)\quad F(X_1, X_2, L, T) \leq 0;$$

$$(7b)\quad H(G, D_1, D_2) \leq 0,$$

tions are also needed without trade: the discreteness of regions prevents the equilibrium condition, $R = 0$, from holding exactly, and it also implies that the "price-taking" and "utility-taking" behavior of regions is in some sense an approximation. Myrna Wooders (1980) formalizes the concept of an approximate equilibrium for a model with one private good and one local public good.

where C_i represents a resident's final consumption of good i, D_i is the quantity of good i used in public good consumption, X_i is the region's total output of good i, and the function F and H describe the production possibility sets for private and public production. Constraint (7) is required for technological feasibility, while (6) is the trade balance constraint, requiring that the total value of net exports be nonnegative. Of course, constraints (6) and (7) are satisfied with equality at the optimum.

Starting from the equilibrium allocation, suppose that a central planner could find a technologically feasible way to reallocate goods and resources so as to raise some individuals' utilities without lowering anyone's utility. To be technologically feasible, the new allocation would have to satisfy (7). Since, however, the original allocation solved the above utility-maximization problem in each region, the new allocation would have to violate constraint (6) in those regions where some residents received a higher utility level. In regions where utilities remained unchanged, constraint (6) would continue to be satisfied with equality. It follows that the total value of net exports, summed over *all* regions, would be negative. This means that the total output of at least one of the two goods would fall short of the total consumption of that good, thereby violating the assumption of technological feasibility. Thus, the equilibrium allocation must have been Pareto efficient.

This is a simple revealed preference argument. It can easily be extended to include a much more general set of assumptions, such as the presence of nontraded goods and congestion effects.[8]

My paper (1987) also presents a model where identical regions engage in commodity trade and pursue different tax and public expenditure policies. However, such differences are wasteful in that model, because they result from the taxation of mobile capital. The present analysis identifies an important efficiency role for differences in tax rates and public expenditure programs across regions with similar types of individuals and production technologies.

IV. Extensions

In this section, I add congestion to the model and show that each region continues to produce only one traded good in equilibrium as long as the public good exhibits scale economies in consumption. Several other extensions to the analysis, and some limitations, are also described.

Congestion is added to the model by assuming that a region's population size is an argument in each resident's utility function: $u=u(Y,G,L,p)$, where Y is the net wage, $w-b$, and $\partial u/\partial L \leq 0$. Two polar cases will serve as useful benchmarks:

$$(8) \qquad u(Y,G,p,L)=u(Y,G,p) \quad \text{for a pure public good;}$$

$$u(Y,G,p,L)=u(Y,G/L,p) \quad \text{for a publicly provided private good.}$$

In the case of a publicly provided private good, each of the L residents can "consume" only G/L units of the region's total public good output, G, because constant returns to scale prevail in consumption. My concern in this section is with the intermediate cases where the public good can be described as "impure."

Two assumptions are imposed on the function u. To state them, I invert u to get $G=G(Y,L,p,u)$. The derivative of this function with respect to L is the marginal rate of substitution between L and G. My first assumption is that this derivative is everywhere nonincreasing with L:

$$(9) \qquad \partial^2 G/\partial L^2 \leq 0.$$

This assumption generalizes the polar cases in (8): $\partial G/\partial L$ always equals zero for a pure public good, and it is independent of L for a publicly provided private good, where $G(Y,L,p,u)$ takes the special form $G=g(Y,p,u)\cdot L$. Although a strict inequality in (9) might be viewed indicating a form of

[8]See my working paper for the more general proof.

scale economies from increased population size, it is not these economies which characterize a pure public good.

For the second assumption, consider the sum of the marginal rates of substitution between G and the head tax, $L \cdot MRS_{Gb} = -L/(\partial G/\partial Y)$. This sum measures the marginal benefits of the public good. While this marginal benefit is independent of L for a publicly provided private good, it necessarily rises with L for a pure public good (where $\partial G/\partial Y$ is independent of L). Such a rise represents the form of scale economies which characterize pure public goods. I shall assume that the public goods in this section exhibit such scale economies:

$$(10) \qquad \partial[-L/(\partial G/\partial Y)]/\partial L > 0.$$

My main conclusion is that each region specializes in producing a single traded good when assumptions (9) and (10) hold. As a result, it is again true that public good supplies differ across regions, although all individuals are identical.

To prove the specialization result, I follow Section I by making b and L control variables in the developer's problem of maximizing R; but I drop G as a control variable, letting b and L determine G through the function $G(w(L)-b, L, p, u)$. The maximization problem may then be stated,

$$(11) \qquad \max_{b,L}\left[r(L)\cdot T + bL - p_G G(w(L)-b, L, p^*, u^*)\right],$$

where p^* and u^* again denote the equilibrium product prices and utility. The first-order condition for b yields the previously derived Samuleson condition:

$$(12) \qquad \partial R/\partial b = (\partial G/\partial Y)[p_G - L/(-\partial G/\partial Y)] = (\partial G/\partial Y)[p_G - L\cdot MRS_{Gb}] = 0.$$

The first-order condition for L is

$$(13) \qquad \partial R/\partial L = [(dr/dL)T - p_G(\partial G/\partial Y)(dw/dL)] + [b - p_G(\partial G/\partial L)] = 0.$$

By (12), the term in the first square brackets in (13) reduces to $(dr/dL)T + (dw/dL)L$. To maintain zero profits in private production, this expression must always equal zero. Thus, the first term in (13) vanishes, leaving $b = p_G(\partial G/\partial L)$. In words, the head tax should equal the "marginal congestion cost," defined as the cost of the additional public good output required to compensate residents for a unit rise in the population.

The specialization result is obtained by showing that, if the first-order conditions hold under incomplete specialization, then the second-order conditions must fail. As argued in Section II, $dw(L)/dL = dr(L)/dL = 0$ over the interval of L's where there is incomplete specialization. Equations (9) and (13) then imply that

$$(14) \qquad \partial^2 R/\partial L^2 = -p_G(\partial^2 G/\partial L^2) \geq 0.$$

By (10) and (12),

$$(15) \qquad \partial^2 R/\partial L \partial b > 0.$$

Thus, the first-order conditions imply that either $\partial^2 R/\partial L^2 > 0$ or $(\partial^2 R/\partial L^2)(\partial^2 R/b^2) - (\partial^2 R/\partial L \partial b)^2 = -(\partial^2 R/\partial L \partial b)^2 < 0$. Each of these cases represents failure of the second-order conditions.

As before, the basic explanation behind the specialization result is that incomplete specialization implies the absence of any diminishing marginal productivity of labor which might serve to offset public good scale economies. If, contrary to my assumptions, these scale economies were exhausted once L reached a level where incomplete specialization occurred, then an equilibrium with incomplete specialization could result.[9] In such an equilibrium, initial scale economies would give each developer the incentive to raise L to the range of values where the equilibrium factor prices provide both industries with zero profits. With the exhaustion of these scale economies, however, there would be no incentive to raise L beyond this range.

[9]Scale economies are "exhausted" if, contrary to (9) and (10), $\partial^2 G/\partial L^2 \geq 0$ and $\partial[-L/(\partial G/\partial Y)]/\partial L \leq 0$.

There are several other extensions to the analysis which I will briefly mention. The conclusion that each region produces only a single traded good clearly does not depend on the number of traded goods in the economy; and it is also unaffected by the presence of nontraded goods, or by the use of capital and labor as direct inputs in public good production. What is important is that different traded goods possess different factor intensities. This assumption is responsible for the crucial observation that, with more than one traded good produced in a region, the residential population can be raised without causing the marginal product of labor to decline.

Complete specialization also occurs when the model is extended to allow individuals to possess different utility functions and different endowments of homogeneous labor. With each developer acting as a "utility taker" for each type of labor, an equilibrium emerges where each region contains only one type of individual and produces only one traded good. Complementarities between different types of labor invalidate these results, but weaker results remain. For the case of a pure public good, Stiglitz (1983, p. 68) demonstrates that all residents which provide the same type of labor in a given region must be identical.[10] In this case, no region ever produces more traded goods than the number of labor types located there, regardless of how many traded goods exist in the entire economy. The basic idea may be simply explained. If a region has more traded goods than labor types, there exists variations in land use which raise the region's total population without altering the fraction of this population providing each type of labor, L^i/L for type i labor. Such changes necessarily raise the marginal benefit of the public good, $\Sigma L^i \cdot MRS^i_{Gb}$, although factor prices stay fixed. The argument then proceeds along the lines of Section II by showing that the ability to vary this marginal product independently of factor prices insures that the net return to land could not have been initially maximized.

The specialization result must again be amended, if more than one immobile factor is included in the model (for example, land and water). If each region's developer seeks to maximize the total value of the region's immobile factors, it is possible to show that each region produces fewer traded goods than the total number of factors located there, mobile and immobile. On the other hand, each region produces at least as many traded goods as the number of immobile factors. Adjustments in the prices of additional immobile factors effectively allow zero profits to be maintained in additional traded good industries. In any case, these results indicate that identical regions still produce different goods in equilibrium, if there are at least as many traded goods as factors in the economy. With more than one immobile factor, however, the assumption that developers maximize the total value of immobile factors is suspect. A less naive political model would recognize the potential for conflict between the owners of different immobile factors.

V. Concluding Remarks

The theory of trade presented here bears some relation to the recent international trade literature on the role of scale economies in production as a basis for trade between similar countries (see Elhanan Helpman and Paul Krugman, 1985). An important distinguishing feature of the present theory is that it is *not* the traded private goods which exhibit scale economies, but rather public goods. Thus, the present theory demonstrates how the characteristics of one production sector can influence the trade of goods produced in another production sector.

An interesting relation also exists between my results and the modern international trade literature on factor mobility. Using the standard Heckscher-Ohlin trade model, Robert Mundel's 1957 classic paper demonstrates that trade and factor mobility are substitutes, both in the "welfare sense" that

[10] Stiglitz assumes that regions do not trade, but his result does not depend on this assumption. For the remainder of this section, I assume that the unit input vectors for all traded goods are linearly independent.

a Pareto-efficient allocation can be obtained by free trade in either goods or factors, and in the "volume-of-trade sense" that an increase in factor trade leads to a reduction in commodity trade. In contrast, a distinguishing characteristic of much of the local public economics literature is that labor mobility and commodity trade can never serve as substitutes, at least in the welfare sense. Labor mobility serves a role which cannot be served by commodity trade alone: namely, it provides a mechanism by which individuals can "vote with their feet" and obtain tax-expenditure packages tailored to their preferences and incomes. In the present paper, labor mobility serves no such role, because all individuals are identical. Yet, the presence of public goods still produces a complementary relation between commodity trade and labor mobility, even if all of the assumptions employed by traditional trade theorists are retained (i.e., two goods, two factors, constant returns to scale in production, identical production technologies across regions, perfect competition, no distortions).[11] To be specific, suppose that all regions initially possess the same number of residents, and labor mobility is not allowed. Since regions are identical in all respects, there is then no incentive to trade goods. If, however, free labor mobility is now allowed, an equilibrium is established where each region necessarily specializes in the production of a single traded good in equilibrium. The equilibrium is Pareto efficient, but only because there exists *both* commodity trade and perfect labor mobility. Commodity trade and labor movements can therefore be viewed as complements in both the volume-of-trade sense and the welfare sense.

[11] James Markusen (1983) provides a thorough analysis of how departures from the traditional assumptions may produce a complementary relation between commodity trade and factor movements.

REFERENCES

Arnott, Richard J. and Stiglitz, Joseph E., "Aggregate Land Rents, Expenditures on Public Goods, and Optimal City Size," *Quarterly Journal of Economics*, November 1979, *93*, 471–500.

Atkinson, Anthony B. and Stiglitz, Joseph E., *Lectures on Public Economics*, New York: McGraw-Hill, 1980.

Berglas, Eitan, "Distribution of Tastes and Skills and the Provision of Local Public Goods," *Journal of Public Economics*, November 1976, *6*, 409–23.

Helpman, Elhanan and Krugman, Paul R., *Market Structure and Foreign Trade*, Cambridge: MIT Press, 1985.

______ **and Pines, David,** "Land and Zoning in an Urban Economy: Further Results," *American Economic Review*, December 1977, *67*, 982–86.

Markusan, James R., "Factor Movements and Commodity Trade as Complements," *Journal of International Economics*, May 1983, *14*, 341–56.

Scotchmer, Suzanne, "Local Public Goods in an Equilibrium: How Pecuniary Externalities Matter," *Regional Science and Urban Economics*, November 1986, *16*, 463–81.

Sonstelie, J. C. and Portney, P. R., "Profit Maximizing Communities and the Theory of Local Public Expenditure," *Journal of Urban Economics*, April 1978, *5*, 263–77.

Stiglitz, Joseph E., "The Theory of Local Public Goods," in Martin S. Feldstein and Robert P. Inman, eds., *The Economics of Public Services*, London: Macmillan, 1977, 274–333.

______, "Public Goods in Open Economies with Heterogeneous Individuals," in Jacques-Francois Thisse and Henry G. Zoller, eds., *Locational Analysis of Public Facilities*, Amsterdam: North-Holland, 1983, 55–78.

Tiebout, Charles M., "A Pure Theory of Local Expenditures," *Journal of Political Economy*, October 1956, *64*, 416–24.

Vickrey, William, "The City as a Firm," in Martin S. Feldstein and Robert P. Inman, eds., *The Economics of Public Services*, London: Macmillan, 1977, 334–43.

Wildasin, David E., "Urban Public Finance," in Jacques Lesourne and Hugo Sonnenschein, eds. *Fundamentals of Pure and*

Applied Economics, Vol. 10, New York: Harwood Academic, 1986.

Wilson, John Douglas, "Trade, Capital Mobility and Tax Competition," *Journal of Political Economy*, forthcoming 1987.

______, "Trade in a Tiebout Economy," working paper, Indiana University, January 1986.

Wooders, Myrna, "The Tiebout Hypothesis: Near Optimality in Local Public Good Economies," *Econometrica*, September 1980, *48*, 1467–85.

Specification Tests of the Lucas-Rapping Model

By MYRA K. HART*

Aggregate fluctuations in employment and unemployment are often explained within a market-clearing framework as intertemporal substitution in labor supply. Under this hypothesis, leisure in the current period is supposed to be highly substitutable with leisure (and goods) in other periods. Consequently, labor supply responds to perceived temporary changes in the real wage although it may be inelastic with respect to permanent changes in the real wage. A very important and influential empirical study of intertemporal substitution was presented by Robert Lucas and Leonard Rapping (1969, L-R). Lucas and Rapping estimated a simultaneous equations model of the aggregate labor market under an adaptive expectations forecasting scheme and found strong support for the intertemporal substitution hypothesis.

Although the adaptive expectations assumption is now rarely used in such models, the intertemporal substitution hypothesis has become quite prominent in equilibrium business cycle theories where it is used to explain how fluctuations in aggregate demand can result in real changes in output, employment and unemployment.[1] Besides the L-R study, relatively few attempts have been made to verify the intertemporal substitution hypothesis, and most have failed to find the kind of intertemporal substitution elasticity estimated by Lucas and Rapping.[2] In addition, the L-R results have never been tested, although Joseph Altonji made some minor data corrections, extended the time period, and reproduced them.[3]

This paper helps to reconcile the L-R results with later findings and also provides a convincing illustration of the importance and usefulness of specification testing. Here the L-R estimates are reproduced and the model is subjected to the kind of overidentification tests now widely used in econometric analysis. The restrictions on the model are easily rejected. This standard specification testing leads to very different conclusions about the empirical importance of the intertemporal substitution hypothesis than those suggested by Lucas and Rapping.

I. The L-R Model and Results

The L-R aggregate labor supply function was derived from a simple household utility-maximization problem involving four commodities: current and future goods consumption, and current and future labor supply. Household labor supply was then expressed as a function of current and expected wages and prices. To estimate an aggregate labor supply function, total man-hours supplied annually (N_t) was deflated by an index of the number of households (M_t) and expressed as a log-linear function of current (W_t) and expected (W_t^*) real wages.

*Departments of Economics, Whittier College, Whittier, CA 90601, and University of Iowa, Iowa City, IA 52240. I am grateful to John F. Kennan for his suggestions and strong encouragement.

[1] In particular see Robert Barro (1976), Lucas (1977), and Robert Hall (1980); more recent examples are Karl Brunner et al. (1983) and George Alogoskoufis (1983). Intertemporal substitution has also been incorporated into recent texts such as Barro (1984) and Thomas Sargent (1979) as part of the new microeconomic foundations of macroeconomics.

[2] Joseph Altonji (1982) briefly reviews the time-series and microdata evidence on intertemporal substitution to date and concludes that the aggregate temporary labor supply elasticity may be between 0.1 and 0.6. This suggests that intertemporal substitution may take place but with an empirical magnitude much smaller than Lucas and Rapping estimated.

[3] Altonji also estimated the L-R model under the assumption of rational expectations using a "compromise procedure" to acquire estimates of the rational expectations forecasts of wages and prices. These results, conditioned on the rational expectations forecasts failed to support the intertemporal substitution hypothesis. Lucas and Rapping (pp. 730–31) justify the adaptive expectations assumption by arguing that adaptive expectations gives good forecasts over their sample period—that adaptive expectations is approximately rational.

The adaptive expectations scheme was postulated for wage and price anticipations, and a Koyck transformation was then used to eliminate W_t and P_t, so that the aggregate labor supply equation estimated by Lucas and Rapping was

$$(1)\quad (n-m)_t=\beta_0+(\beta_1-\lambda\beta_2)w_t - (1-\lambda)\beta_1 w_{t-1}+(1-\lambda)\beta_3(p_t-p_{t-1}) + (1-\lambda)(n-m)_{t-1},$$

where λ is the speed of adjustment in the expectations equations, and lowercase letters are used to represent the natural logs of the original data series.

Lucas and Rapping also derived and estimated an aggregate marginal productivity condition for labor from an aggregate production function with constant elasticity of substitution. Dynamics were imposed by assuming partial adjustment of output and employment. Let y_t be the natural log of real GNP, q_t an index of labor quality, and w_t, p_t, and n_t as defined above. Then the marginal productivity condition estimated by Lucas and Rapping was

$$(2)\quad (n+q-y)_t=c_0+c_1(w-q)_t + c_4(n+q-y)_{t-1}+(c_2-1)(y_t-y_{t-1}).$$

The marginal productivity condition and the aggregate labor supply equation were estimated by L-R using two-stage least squares and aggregate U.S. time-series for the years 1930–65. The structural estimates obtained seemed quite reasonable in light of the many sign predictions given by the theory underlying the model. Of particular interest were the estimates of labor supply elasticities with respect to the real wage. The coefficient on w_t in the supply equation gave an estimate of the short-run elasticity of labor supply. Lucas and Rapping reported a significantly positive value of 1.40 for this short-run elasticity. The long-run elasticity estimate was 0.03, or essentially zero. Thus the intertemporal substitution hypothesis seemed to be substantiated by these findings.

II. Specification Tests

One way to check the specification of a simultaneous equations model is to test the overidentifying restrictions on each equation. These overidentifying restrictions imply a set of cross-equation restrictions on the reduced form of the model, which can be tested with either a likelihood ratio test or a Wald test (these two tests are asymptotically equivalent). Rejection of the reduced-form restrictions by either of these tests indicates a misspecified model. In this section the reduced-form restrictions on the L-R model are derived and tested with both Wald and likelihood ratio tests.

Aside from the usual normalizations, there are five structural form constraints implied by the marginal productivity condition (1) and four constraints derived from the specification of the aggregate labor supply equation (2). All but two of these constraints are simple exclusion restrictions. To state the constraints specifically, rewrite (1) and (2) as $\Gamma y_t = Bz_t + u_t$, where $u_t \sim NID(0,\Omega)$, $y_t=[(n-m)_t, w_t]'$, and $z_t=[1, w_{t-1}, p_t-p_{t-1}, (y-m)_t, q_t, (q+n-y)_{t-1}, y_t-y_{t-1}, (n-m)_{t-1}]'$. The specification of the L-R model then requires that

$$(3a)\quad \Gamma=\begin{pmatrix}\gamma_{11} & \gamma_{12}\\ \gamma_{21} & \gamma_{22}\end{pmatrix}=\begin{pmatrix}1 & c_1\\ 1 & -(\beta_1-\beta_2)\lambda\end{pmatrix};$$

$$(3b)\quad B=\begin{pmatrix}b_{10} & b_{11} & b_{12} & b_{13} & b_{14} & b_{15} & b_{16} & b_{17}\\ b_{20} & b_{21} & b_{22} & b_{23} & b_{24} & b_{25} & b_{26} & b_{27}\end{pmatrix}$$

$$=\begin{pmatrix}c_0 & 0 & 0 & 1 & c_1-1 & c_4 & c_2 & 0\\ \beta_0' & -(1-\lambda)\beta_1 & (1-\lambda)\beta_3 & 0 & 0 & 0 & 0 & (1-\lambda)\end{pmatrix}.$$

Now the constraints on the structural form can be written

$(MP1)\, b_{11}=0$ $\quad$ $(S1)\, b_{23}=0$

$(MP2)\, b_{12}=0$ $\quad$ $(S2)\, b_{24}=0$

$(MP3)\, b_{17}=0$ $\quad$ $(S3)\, b_{25}=0$

$(MP4)\, b_{13}=1$ $\quad$ $(S4)\, b_{26}=0$

$(MP5)\, \gamma_{12}-b_{14}=1$

The cross-equation restrictions on the reduced form implied by these structural form constraints were derived explicitly following R. P. Byron (1974). Writing the reduced form as

$$(4)\quad y_t = \begin{pmatrix} \pi_{10} & \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} & \pi_{15} & \pi_{16} & \pi_{17} \\ \pi_{20} & \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24} & \pi_{25} & \pi_{26} & \pi_{27} \end{pmatrix} \times Z_t,$$

the constraints derived from the overidentifying restrictions on the marginal productivity condition are

$$(5a)\quad \pi_{11}/\pi_{21} = \pi_{12}/\pi_{22} = \pi_{17}/\pi_{27} = (\pi_{13}+1)/\pi_{23} = (\pi_{14}-1)/(\pi_{24}+1) = -c_1.$$

The reduced-form constraints derived from the supply equation are

$$(5b)\quad \pi_{13}/\pi_{23} = \pi_{14}/\pi_{24} = \pi_{15}/\pi_{25} = \pi_{16}/\pi_{26} = \beta_1 - \beta_2\lambda.$$

The standard test of the reduced-form restrictions is a likelihood ratio test which requires maximum likelihood estimation of both the constrained and unconstrained reduced forms.[4] The likelihood ratio statistic is $L = T[\log|\tilde{\Sigma}_0| - \log|\tilde{\Sigma}|]$ where $\tilde{\Sigma}_0$ and $\tilde{\Sigma}$ are the maximum likelihood estimates of the covariance matrices of the restricted and unrestricted reduced-form residuals. The estimate $\tilde{\Sigma}$ was calculated from residuals of OLS regressions of each unrestricted reduced-form equation while $\tilde{\Sigma}_0$ was obtained from the residuals of three-stage least squares estimation of the structural form system.

The value of L calculated for the L-R model was 144.7. For a correctly specified model L would be asymptotically *chi*-squared with 5 degrees of freedom. Thus the likelihood ratio test rejects the overidentifying restrictions on the L-R model at any reasonable level of significance.

Byron has also suggested a Wald test of the reduced-form restrictions which under the null is asymptotically equivalent to the likelihood ratio test described above. Wald tests are carried out using estimates of the unrestricted reduced-form and linear approximations of the reduced-form constraints. Express the reduced-form constraints as

$$(6)\quad h(\Pi) = \begin{pmatrix} \pi_{11}/\pi_{21} - \pi_{12}/\pi_{22} \\ \pi_{11}/\pi_{21} - \pi_{17}/\pi_{27} \\ \pi_{11}/\pi_{21} - (\pi_{13}+1)/\pi_{23} \\ \pi_{11}/\pi_{21} - (\pi_{14}-1)/(\pi_{24}+1) \\ \pi_{13}/\pi_{23} - \pi_{14}/\pi_{24} \\ \pi_{13}/\pi_{23} - \pi_{15}/\pi_{25} \\ \pi_{13}/\pi_{23} - \pi_{16}/\pi_{26} \end{pmatrix} = 0.$$

The restrictions are rejected if $h(\tilde{\Pi})$ is far enough from zero where $\tilde{\Pi}$ is the vector of unrestricted maximum likelihood estimates of the reduced-form parameters. Writing the reduced form as $Y = Z\Pi + V$ where Y is the $T\times 2$ matrix of endogenous variables, Z is the $T\times k$ matrix of exogenous variables, Π is $k\times 2$ and V is $T\times 2$, the Wald statistic then becomes

$$(7)\quad W = h(\tilde{\Pi})'\left\{H\left[(1/(T-k))\tilde{V}'\tilde{V} \times (Z'Z)^{-1}\right]H'\right\}h(\tilde{\Pi}),$$

where $H' = dh(\tilde{\Pi})/d\Pi$, $\tilde{\Pi}$ is the vector of unrestricted OLS reduced-form parameter estimates and $\tilde{V}$ is the $T\times 2$ matrix of unrestricted OLS residuals.

The statistic W is asymptotically *chi*-squared with degrees of freedom determined by the number of restrictions on the reduced form. For the L-R model, W has seven degrees of freedom and was calculated to be 64.92. Thus, this Wald test readily confirms the likelihood ratio test result.

III. Conclusion

This paper illustrates that "standard" specification testing would have rejected the L-R specification and lead to very different conclusions about the empirical importance

[4] See A. C. Harvey (1981), for example.

of the intertemporal substitution hypothesis. The significantly positive short-run and zero long-run labor supply elasticities reported by Lucas and Rapping were estimated from a misspecified model. In view of the fact that later studies have not confirmed the sign or magnitude of L-R's substitution elasticity, this is a very useful example of the importance of specification testing. Tests which are standard today would immediately have indicated the fragility of the L-R results and possibly changed the direction of macroeconomic research.

REFERENCES

Alogoskoufis, George S., "The Labor Market in an Equilibrium Business Cycle Model," *Journal of Monetary Economics*, January 1983, *11*, 117–128.

Altonji, Joseph G., "The Intertemporal Substitution Model of Labor Market Fluctuations: An Empirical Analysis," *Review of Economic Studies*, Special Issue, 1982, *49*, 783–824.

Barro, Robert J., "Rational Expectations and the Role of Monetary Policy," *Journal of Monetary Economics*, January 1976, *2*, 1–32.

______, *Macroeconomics*, New York: Wiley & Sons, 1984.

Brunner, Karl, Cukierman, Alex and Meltzer, Allan H., "Money and Economic Activity, Inventories and Business Cycles," *Journal of Monetary Economics*, May 1983, *11*, 281–319.

Byron, R. P., "Testing Structural Specification using the Unrestricted Reduced Form," *Econometrica*, September 1974, *42*, 869–83.

Hall, Robert E., "Labor Supply and Aggregate Fluctuations," in K. Brunner and A. Meltzer, eds., *Carnegie Rochester Conference Series on Public Policy: On the State of Macro-Economics*, Spring 1980, *12*, 7–33.

Harvey, A. C., *The Econometric Analysis of Time Series*, New York: Halsted Press, 1981.

Lucas, Robert E., Jr., "Understanding Business Cycles," in K. Brunner and A. Meltzer, eds, *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conference Series on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 7–29.

______ **and Rapping, Leonard A.,** "Real Wages, Employment, and Inflation," *Journal of Political Economy*, September/October, 1969, *77*, 721–54.

Sargent, Thomas J., *Macroeconomic Theory*, New York: Academic Press, 1979.

Informal Job Search and Black Youth Unemployment

By HARRY J. HOLZER*

One potential source of low employment for young blacks which has long been suggested but rarely analyzed empirically is the network of contacts available to them. There are many reasons for believing that blacks may enjoy fewer benefits from such contacts than do whites. These include the rise in the number of welfare-dependent black households with no employed members, the high unemployment rates of older black males (and their low representation in skilled blue-collar positions), and lower confidence in the recommendations of employed blacks by white employers.

Contacts through friends and relatives can be considered part of a general category of informal job search, which also includes direct applications to firms from walk-ins without referral. More formal methods of search include state or private employment agencies, responding to newspaper ads, school or community placement services, and other institutional activities (see Albert Rees, 1966). Blacks might do relatively better with formal methods than with informal ones because informal methods involve fewer explicit or objective criteria by which to judge applicants, and instead rely heavily on subjective judgments by employers or references. This is particularly true for direct applications from walk-ins, where the applicant's race might be among his or her most salient features.[1]

In this paper, I use data from the 1981 and 1982 panels of the *National Longitudinal Survey of Youth* (*NLS*) to test for racial differences in the use and effectiveness of various job search methods. I also decompose the total observed difference in employment probabilities into components attributed to each method of search; and further into differences in use, job offers, and job acceptances based on all methods.

I. Summary Results

The 1981 panel of the *NLS* asked a broad range of questions regarding job search activities in the previous month. Anyone who had searched for work was asked whether he or she had used any of about thirteen different methods. If so, the respondent was then asked whether that method resulted in a job offer and whether the offer had been accepted.

The sample used in the analysis below is limited to white and black males who are not currently enrolled in school or enlisted in the military. The ages of the respondents range from 16 to 23. The sample includes all individuals who searched in the previous month and are currently unemployed, or who are currently employed but whose duration of employment is thirty days or less.

The mean probabilities of becoming employed through each of the search methods (including a composite category) appear in Table 1. Probabilities are also listed for each method for use, offers (conditional on use), and acceptances (conditional on offers). All means are weighted by sample weights, to correct for the oversampling of low-income whites in the *NLS*.

The results show that, for both white and black youths, the most frequently used methods of search are checking with friends and relatives, and direct application without referrals. These are also the two most productive methods, in terms of offers and accep-

*Michigan State University, East Lansing, MI 48824–1038. I thank Dan Hamermesh and John Bound for helpful comments, as well as seminar participants at Michigan, Michigan State, and the NBER. Eric Krupka provided excellent research assistance. Funding was provided by NSF grant no. SES-8408876.

[1]The higher use of formal methods by blacks and informal methods by whites has been noted previously (see, for instance, Thomas Bradshaw, 1973). Differences in relative effectiveness between the two groups has been suggested by several authors but never demonstrated.

TABLE 1—SEARCH METHOD USE AND OUTCOMES, 1981 *NLS*, UNEMPLOYED WHITES (W) AND BLACKS (B)

	Friends/Rels.		Direct App.		State Agency		Newspaper		Other	
	W	B	W	B	W	B	W	B	W	B
Percent Using each Method in Previous Month:	.862	.818	.808	.758	.548	.501	.584	.556	.528	.509
Percent of Users Obtaining Offers:	.183	.156	.211	.094	.092	.079	.109	.065	.169	.130
Percent of Offers Accepted:	.809	.801	.654	.649	.511	.633	.339	.800	.627	.546
Probability of Gaining Employment:	.128	.102	.111	.049	.026	.025	.022	.029	.056	.036

Note: Samples include nonenrolled and nonenlisted males who were unemployed and searching for work one month prior to the survey. Sample sizes are 398 for whites and 211 for blacks.

tances generated. In fact, the two informal methods account for almost 70 percent of jobs obtained by whites and almost 60 percent of those obtained by blacks.

As for comparisons between blacks and whites, the frequency of search method use is a bit lower for blacks in each category. However, most of these differences are not significant.[2] Much more substantial are the differences in the probabilities of obtaining offers between the two groups. For each method, whites have higher probabilities of obtaining offers than do blacks.[3] The largest difference, in both absolute and percentage terms, occurs for the method of direct application. Racial differences in the conditional probabilities of accepting job offers also appear to be smaller and more mixed than those for the probabilities of receiving offers. Finally, the employment probabilities of the last row in Table 1 show that racial differences are largest for the two informal methods.

The questions in the 1982 panel of the *NLS* dealt with search methods used by the respondent to obtain his most recent job. All those who listed a job in the previous year were asked whether they were already employed when they obtained it; whether they had been searching for work when they obtained it; if they had, what methods they used and how many weeks they spent searching; and which method resulted in obtaining the job.

Summary results from these data appear in Table 2. For each of the five search methods, we find the percentages of job holders and job seekers who obtained their jobs in the previous year through this method. Monthly probabilities for those obtaining and seeking jobs through each method (calculated from number of weeks spent searching),[4] and the percent of all job seekers using each method, are also included. As in Table 1, all means are weighted and the sample is restricted to all unemployed searchers who are neither enrolled nor enlisted.[5]

[2]Standard errors on the means of the use probabilities are in the vicinity of .01–.02 for whites and .02–.03 for blacks on most methods. Since the samples of whites and blacks are independent, the standard errors on differences in means are calculated by $SE = ((S.E.)_w^2 + (S.E.)_B^2)^{1/2}$. Thus, standard errors on differences in search methods use are approximately .02–.04.

[3]Standard errors on the conditional offer probabilities are approximately .03 for whites and .04 for blacks, implying standard errors of about .05 for the differences by the formula stated in fn 2.

[4]Assuming constant transition probabilities, the monthly probabilities for those obtaining jobs through method j can be approximated by $4/w_j$ where w_j reflects weeks spent searching.

[5]Due to the sequencing of questions in the 1982 *Survey*, we have no data on the job-seeking activities of those without employment in the previous year. Since we are considering nonenrolled and nonenlisted young males, I make the assumption that all such individuals spent some time searching for work in the previous year. However, those individuals who claim to have *not* been searching when they obtained their most recent jobs are omitted from the sample.

TABLE 2—METHOD OF OBTAINING JOB IN PREVIOUS YEAR, 1982 *NLS*, UNEMPLOYED WHITES AND BLACKS

	Friends/Rels.		Direct App.		State Agency		Newspaper		Other	
	W	B	W	B	W	B	W	B	W	B
Percent of Job Holders Who Obtained Job through:	.352	.329	.337	.257	.032	.048	.067	.036	.212	.330
Percent of Job Seekers Who Obtained Job through:	.318	.255	.304	.199	.029	.037	.060	.028	.191	.256
Monthly Probability of Employment for Job Holders through:	.460	.332	.505	.412	.436	.428	.474	.563	.464	.326
Monthly Probability of Employment for Job Seekers through:	.146	.085	.154	.082	.013	.016	.028	.016	.089	.083
Percent of Job Seekers using each Method:	.691	.669	.598	.481	.313	.418	.351	.335	.336	.391

Note: Samples include nonenrolled and nonenlisted males who were unemployed and searching when they obtained their most recent (rows 1 and 3) jobs plus those who held no employment in the previous year (rows 2, 4, and 5). Sample sizes are 1269 and 472, respectively, for whites and blacks who had jobs, and 1405 and 609 for the total sample.

The results show, as before, that friends and relatives and direct applications produce the largest number of accepted jobs for both whites and blacks. Also as in Table 1, the largest racial differentials exist for direct applications. However, the racial differentials for probabilities based on friends and relatives are larger than they were in Table 1, especially when conditional monthly probabilities are included. It therefore appears from this table that both informal search methods are important determinants of differences in employment probabilities between young whites and blacks.

II. Decomposition of Employment Differences

The summary data of Tables 1 and 2 can be used to decompose the overall racial difference in employment probabilities into fractions accounted for by each search method. The fractions of this difference which are accounted for by differences in search method use, offer probabilities and acceptance probabilities can also be calculated.

The decomposition of the overall difference in probability of employment for the 1981 data appears in Table 3. The first row lists the absolute differences in overall employment probabilities based on Table 1. All other numbers in the table reflect fractions of these differences. The second row represents the fractions of the overall difference which are accounted for by differences in employment probabilities for each method of search.

In order to further decompose these percentages into components based on use, offers, and acceptance, I take logs of the ratios of employment probabilities for whites and blacks:

$$(1)\quad \ln\left(P(E_j)^W/P(E_j)^B\right)$$
$$\equiv \ln\left(\frac{P(Use_j)^W}{P(Use_j)^B}\right)+\ln\left(\frac{P(Off_j|Use_j)^W}{P(Off_j|Use_j)^B}\right)+\ln\left(\frac{P(Acc_j|Off_j)^W}{P(Acc_j|Off_j)^B}\right)$$
$$\equiv \Delta\ln P(Use_j)+\Delta\ln P(Off_j|Use_j)+\Delta\ln P(Acc_j|Off_j),$$

TABLE 3—DECOMPOSITION OF RACIAL DIFFERENCES IN PROBABILITIES OF GAINING EMPLOYMENT, 1981 *NLS*

	Total	Friends/ Relatives	Direct Application	State Agency	Newspaper	Other
Total Difference in Employment Probability in Previous Month:	.104	.025	.065	.001	−.007	.020
Percentage of Difference Due to each Method:	1.000	.244	.628	.007	−.070	.192
Percentage Due to Differences in						
Use of Methods:	.154	.058	.046	.022	.012	.016
Receiving Offers:	1.028	.176	.576	.037	.125	.114
Accepting Offers:	−.181	.010	.006	−.052	−.207	.062

Note: Calculations, based on data from Table 1, are described in the text.

where the probabilities correspond to those included for each method of j in Table 1. Dividing each of the three components by the overall log ratio gives us the percentage of each difference in employment probabilities accounted for by use, offers, and acceptances. When, in turn, these percentages are multiplied by the respective $\Delta P(E_j)/\Delta P(E)$ percentages of the second row, we get the percentages of the total difference in employment probabilities accounted for by use, offers, and acceptances with each method. These last numbers appear in the third, fourth, and fifth rows, respectively, of Table 3. Given these calculations, we can add across each row to obtain the percentages of the total difference accounted for by use, offers, and acceptances from all search methods. These numbers appear in the first column of Table 3. Likewise, we can add across each column to obtain the percentages of the total difference due to each method, which appear in the second row.

The results of Table 3 show that the two informal search methods account for about 87 percent of the total racial difference in employment probabilities. Direct applications alone account for almost 63 percent. The percentage of the total difference accounted for by friends and relatives is second in magnitude, while the composite category of "other methods" picks up the rest.

Furthermore, differences in offer probabilities for the five methods together can explain the entire racial difference in youth employment probabilities. Though differences in frequency of search method use could explain an additional 15 percent of the total difference, this component is fully counteracted by the higher conditional probabilities of accepting offers among young blacks.

Table 4 presents a decomposition of the racial differences in employment probabilities based on data from the 1982 panel. The first two rows are comparable to those of Table 3, in presenting absolute and percentage differences in monthly employment probabilities attributable to each method of search. The third and fourth rows decompose the percentage differences of the second row into portions attributable to annual and conditional monthly differences for each method, based on the following equation:

$$\text{(2)}\quad \ln\left(P(E_{jm})^{w}/P(E_{jm})^{B}\right)$$
$$\equiv \ln\left(\frac{P(E_j)^{w}}{P(E_j)^{B}}\right)+\ln\left(\frac{P(E_{jm}|E_j)^{w}}{P(E_{jm}|E_j)^{B}}\right)$$
$$\equiv \Delta\ln P(E_j)+\Delta\ln P(E_{jm}|E_j),$$

where $P(E_{jm})$ reflects the monthly probabilities for job seekers, $P(E_j)$ reflects the annual ones, and $P(E_{jm}|E_j)$ reflects the monthly probabilities for job holders using each method j in Table 2. Dividing each of these terms by the sum and multiplying by the percentage differences in the second row give us the percentages due to annual and monthly probabilities of the third and fourth rows of Table 4.

TABLE 4—DECOMPOSITION OF RACIAL DIFFERENCES IN PROBABILITIES OF GAINING EMPLOYMENT, 1982 *NLS*

	Total	Friends/ Relatives	Direct Application	State Agency	Newspaper	Other
Total Difference in Monthly Employment Probability:	.148	.061	.072	−.003	.012	.006
Percentage of Total Difference Due to each Method:	1.000	.412	.486	−.020	.081	.041
Probability of Holding a Job During Previous Year:	.443	.168	.327	.003	.110	−.165
Probability of Obtaining this Job Within a Month:	.557	.244	.159	−.023	−.029	.206
Percent of Total Difference Due to						
Use of Methods:	.081	.024	.168	−.028	.007	−.090
Receiving/Accepting Offers:	.919	.388	.318	.008	.074	.131

Note: Calculations, based on data from Table 2, are described in the text.

Finally, the last two rows decompose the percentage differences in the second row for each method into components attributable to use and to receiving and accepting offers from each method. These are based on equations similar to those in equation (1) except that the final row here reflects a residual difference in monthly probabilities after differences in use have been accounted for.[6]

The results of Table 4 show that friends and relatives and direct applications can account for about 90 percent of the difference in monthly employment probabilities between young blacks and whites. Direct application accounts for almost half of the differential while friends and relatives account for over 40 percent. Both contribute substantially to differentials at the annual and monthly levels.

Furthermore, the results show that differences in use account for about 8 percent of the total monthly difference. While differences in direct applications account for almost 17 percent of the total differential, part of this difference is overturned by higher use of other methods by blacks. As for the difference between receipt and acceptance of offers, evidence from an additional question in the 1982 panel shows a much higher rate of offer rejections among whites than among blacks.[7] Thus virtually the entire difference in employment probabilities is accounted for by differences in probabilities of receiving offers, as in the 1981 *NLS*.[8]

In order to relate these calculations to black youth *un*employment, I use the fact the employment probabilities reflect the inverse of unemployment durations (assuming constant transition probabilities); and then we must consider the extent to which differences in unemployment between young blacks and whites reflect differences in duration as opposed to frequency in that state. Calculations on the *NLS* data which I report

[6]Due to the omission of data on job-seeking activities of those without jobs in the previous year, the calculations performed here assume that this group of individuals used search methods in similar proportions to those with jobs in the previous year.

[7]In response to the question, "Did you reject any offers achieved through any of these methods?", 21.0 percent of white jobholders and 11.8 percent of blacks answered yes. The higher rate of job rejection (conditional on having held a job) among whites approximates that observed in the 1981 panel.

[8]These findings contrast somewhat with those found by John Barron and Otis Gilley (1981). Using a special supplement on job search in the *CPS* of May 1976, they found that method of search had little effect on employment probabilities once reservation wages and overall search intensity were controlled for. Racial differences in employment probabilities do appear in their work after including these controls, though smaller in magnitude (.07 in their work, .10 here). These differences may at least partly reflect the focus of this study on youth and on males only.

elsewhere (1986a) show that differences in duration account for about 65–80 percent of the total unemployment differences between black and white youth. Therefore the differences in employment probabilities generated by informal search methods would explain about 57–72 percent of the overall difference in unemployment between young blacks and whites. Since differences in offer probabilities can account for the entire difference in duration, they would also account for 65–80 percent of the total difference in unemployment between the two groups.

It should be noted that differences in offer probabilities between blacks and whites might still reflect differences in search choices. For instance, it is possible that reservation wages affect the decisions of individuals to seek explicit offers from particular firms as well as the decisions of whether to accept such offers. Other evidence from the *NLS* suggests that the reservation wages of young blacks are higher relative to offered wages than are those of young whites, and that these differences partly explain the longer unemployment durations of the blacks.[9]

The data on time spent searching in the 1981 *NLS* also indicate that young whites appear to spend less time on formal methods and more time on informal methods than do young blacks.[10] These choices are, in fact, consistent with a model in which individuals choose search intensities for each method on the basis of relative productivities and costs for each.[11] Thus, young blacks may choose lower search intensity than whites when using informal methods as a response to lower offer probabilities and/or lower wage offers for any given intensity. However, differences between the survey questions used to gauge search intensities and those used for outcomes made any attempts to estimate the effects of the former on the latter for whites and blacks inconclusive here.[12]

Finally, it should be noted that equations for offer probabilities have been estimated in which I use various personal and household characteristics (including the presence and employment status of household members) as explanatory variables. However, their ability to explain observed racial difference was quite mixed and often limited.[13] It is therefore possible that the problems faced by young blacks in their use of informal search methods afflict a wide range of individuals, regardless of personal and background factors.

III. Conclusion

The results of this paper show that the two informal methods of search, especially direct application without reference, account for 87–90 percent of the difference in youth employment probabilities between blacks and whites. Furthermore, virtually all of this reflects differences in the ability of these methods to produce job offers, as opposed to differences in methods used or job accep-

[9]See my earlier paper (1986b). Results of that paper, based on data from the 1979 and 1980 panels of the *NLS*, showed that the ratio of reservation wages to received wages was about 15 percent higher for unemployed blacks, although the levels of reservation wages were comparable for the two groups. The ratio of reservation to received wages in the 1981 *NLS* was 9.1 percent higher for blacks, while in the 1982 panel it was 6.0 percent higher among the unemployed.

[10]For those using each method, time spent (in minutes per previous week) for whites and blacks, respectively, are 322.3 and 210.6 for friends/relatives; 397.3 and 252.2 for direct application; 186.6 and 292.5 for state agencies; 223.6 and 292.2 for newspapers; and 205.4 and 266.0 for other methods.

[11]See my paper (1986c) for such a model.

[12]Estimated equations for whether or not individuals had received offers in the previous month from using either informal method show virtually no effect from the inclusion of these time intensities on racial differences. However, since the search intensity questions refer only to the previous *week* while outcome questions refer to the previous *month*, many individuals who had received and accepted offers earlier in the month did not answer the time-intensity questions. The omission of these data should certainly bias the estimated effects of time-intensities toward zero.

[13]For more detail on results from these estimated equations, see my paper (1986a).

tance rates. The evidence thus strongly suggests that young blacks face more severe barriers when using informal rather than formal search methods, possibly because of the greater role played by personal contacts and subjective employers' impressions in the former. Certain search choices of these youth (such as search intensities for each method used and reservation wages) may also play some role in generating these outcomes.

These findings suggest some potentially important implications for policy approaches on the black youth employment issue. Those approaches which stress formal, institutional mechanisms for job placement may be less successful, while those which provide lessons in informal job search and direct application procedures may be more successful for young blacks. But disadvantages in the network of friends and relatives facing blacks are more difficult to overcome through policies of either type. More research on the causes of disadvantages for blacks who use informal search methods is necessary before remedies for this problem can be promoted with confidence.

REFERENCES

Barron, John and Gilley, Otis, "Job Search and Vacancy Contacts: Note," *American Economic Review*, September 1981, *71*, 747–52.

Bradshaw, Thomas, "Jobseeking Methods Used by Unemployed Workers," *Monthly Labor Review*, February 1973, *96*, 35–40.

Holzer, Harry J., (1986a), "Informal Job Search and Black Youth Unemployment," NBER Working Paper No. 1860, March 1986.

______, (1986b), "Reservation Wages and their Labor Market Effects for White and Black Male Youth," *Journal of Human Resources*, Spring 1986, *21*.

______, (1986c), "Search Method Use by Unemployed Youth," NBER Working Paper No. 1859, March 1986.

Rees, Albert, "Information Networks in Labor Markets." *American Economic Review Proceedings*, May 1966, *56*, 559–66.

Center for Human Resource Research, *National Longitudinal Survey of Youth*, Ohio State University, 1981; 1982.

Licensing and Nontransferable Rents

By JOHN R. LOTT, JR.*

Traditionally, restrictive licensing is assumed to create monopoly profits by restricting output, and therefore to produce two kinds of social costs: the deadweight loss due to reduced output and the resources devoted to rent seeking. However, the fact that nonsalvagable resources spent on rent seeking create their own barriers to entry has not been recognized. By increasing nontransferable rents, licensing prevents the least costly producers from entering, and thus produces a third kind of social cost. While Harold Demsetz' (1982) dismissal of the traditional notion of entry barriers is correct when assets are transferable, the idea of entry barriers is still useful when assets are nontransferable, as this note shows in the case of professional licensing.[1]

I. Resalable Property Rights

Suppose a profession's (for example, barbering's) demand curve (D) and marginal cost curve (S) are as shown in Figure 1, and the government restricts the number of barbers by issuing only Q^* licenses. New barbers can therefore only enter when old barbers leave the profession. The government's aim could be to maintain the price at P_m.[2]

Risk-neutral individuals are willing to pay the present value of the monopoly rents for a license. Since those with the lowest opportunity costs would receive the highest rents, I assume that they will obtain the licenses initially. Individual opportunity costs may change over time so that the marginal cost curve of the initial licensees shifts up to S'.[3] If potential entrants are represented by the old supply curve (S) and the license is transferable, incumbents would sell their licenses to the lower-cost entrants, and Q^* would continue to be produced at the lowest possible resource cost. Without transferability, however, existing barbers would continue producing as long as they received positive rents (as they do with S'), thus causing a social loss of *abcd*. This loss is in addition to the deadweight loss from reduced output, *cef*, and the loss due to rent seeking. Hence, if licenses are nontransferable, the traditional idea of barriers to entry is relevant—existing licensees have an advantage over potential newcomers simply because of past investments that they have made in obtaining the license. Transferability is crucial to Demsetz' dismissal of the traditional definition of entry barriers, since opportun-

*Hoover Institution, Stanford, CA 94305, and Department of Agricultural Economics, and Economics, Montana State University, Bozeman, MT 59717. I thank Edgar Browning, Paul Evans, Gertrud Fremling, Morgan Reynolds, and an anonymous referee for helpful comments. I also thank the Political Economy Research Center for support.

[1]Demsetz' argument is written entirely in terms of barriers to entry for firms and not individuals. His analysis deals with the meaning of barriers to entry in industrial organization. In that context, licensure is analogous to trademark, copyright, and patent restrictions (p. 56). As he points out (pp. 47–49), since costs are opportunity costs, the traditional concept of barriers to entry in industrial organization—as certain firms having cost advantages due to the ownership of specific resources—is meaningless. Even if specific assets are not resalable between firms, their use still represents a cost to the extent that *firms are ultimately resalable*. I shall argue that while the traditional approach is usually inappropriate in analyzing firm behavior, it does have unrecognized implications for individuals since *individuals are not salable*. This distinction between firms and individuals extends beyond licensing. For instance, an individual's "goodwill," unlike a firm's brand name, can serve as a barrier to entry because of its nonsalability (see my 1986 and 1987 papers).

[2]Alternatively, it could make new entrants pay a fixed cost equal to $P_m - P_o$, and allow Q^* to vary over time. While this would allow *some* of the new lowest-cost producers to enter, nonsalability of existing licenses still causes barriers to entry.

[3]Opportunity costs may change systematically over time as some individuals lose certain abilities with age—not following innovations, decreasing strength or speed. Even though the present value of someone's quasi rents from the license is greater than anyone else's at present, it might no longer be true ten years from now.

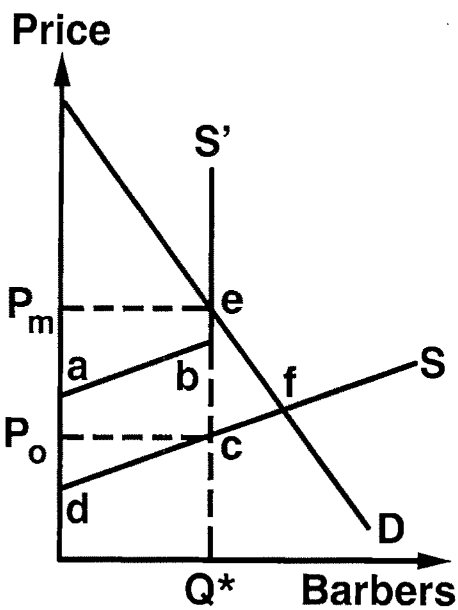


FIGURE 1. THE THIRD SOCIAL COST OF NONTRANSFERABLE PROFESSIONAL LICENSING

ity costs (the foregone sale of licenses) prevent the initial licensees from possessing a differential cost advantage in competing against newcomers. Without transferability, the ownership of specific resources thus produces cost advantages.

Professional licensing causes rents to be competed away through queuing and sometimes through additional investments in human capital to obtain nontransferable rights. Each new entrant incurs rent-seeking costs, with no payments to purchase the slots of existing licensees. Many professions require minimum times of study before allowing the state board examinations to be taken.[4] Obviously, such a time restriction increases the cost of entering the profession. While licensees receive quasi rents from these past investments, they cannot sell the corresponding rental stream. Since the costs of obtaining a license are sunk, even if the costs of existing licensees rise (as shown in Figure 1), these individuals will remain in the market as long as the rents that they receive are positive. Even if potential entrants have a cost curve represented by S, nontransferability will make it in the interest of existing licensees to leave the market only when their costs rise above P_m. Thus, existing licensees may continue despite superior potential entrants.

II. Why Are Licenses Nontransferable?

I still must explain why licenses are not resalable. Since incumbents would gain from selling the licenses to lower-cost producers, it seems that incumbents should favor such an option. But they generally do not. One explanation for nontransferability is that licensing benefits society—consumers and high-quality producers—by preventing quality from being "underproduced." Hayne Leland (1979) argues that low-quality people in a profession will drive out high-quality ones, by lowering the price everyone receives below high-quality people's opportunity costs. Licensing can therefore be a method of preventing low-quality people from entering. However, licensing solves this problem only if it is nontransferable.

When the customers of a licensed service are unable to differentiate low from high-quality providers and when the earnings of these providers in other occupations are positively correlated with their abilities in the regulated profession, transferability of licenses would result in the lowest-quality entrant offering the most to obtain a license. Leland (p. 1335) finds that, in the extreme, the average quality produced can be so low that the correspondingly low fee makes high-quality producers leave the market and "complete market degeneration" can occur, with their (transferable) licenses being worthless. Although licensing combined with nonresalability can ensure the quality of new entrants, competition to enter based on the quality at the moment of entry, like competition based on time, creates nontransferable

[4]Examples of the minimum length of study at certified schools required for professional licenses in Texas are barbers and cosmetologists, 1500 hours of classes in not less than 9 months; dental hygienists, 2 terms of 8 months each of instruction; vocational nurse, not less than 2 years; optometrists, 6 terms of 8 months each; and podiatrists, 4 terms of 8 months each (Texas State O.I.C.C., 1983).

sunk investments and thus creates its own barriers to entry. Incumbents will remain in business because of the quasi rents even when potential entrants offer higher quality.

A more traditional public choice explanation would be that licensing is being used to capture monopoly rents and that professions may find it difficult convincing legislatures to set specific limits on the number of doctors, barbers, etc. As a substitute, professions push for restrictions based on "plausible" public interest arguments. For instance, a minimum length of schooling raises entry costs and limits the size of the profession, while criticism can be countered by the necessity to preserve quality (see Reuben Kessel, 1958, p. 29). Asking for unrestricted resalability would contradict the quality argument and hence may jeopardize any legislated restrictions.

Actual restrictions normally specify time rather than number of classes (see fn. 4). However, if increased quality is the sole objective, the latter constraint should suffice. Even if state board exams imperfectly measure proficiency, time constraints cannot be explained solely by using time as a proxy for hard-to-test experience; it might explain the 1500 hour provision for barbering, but not the nine-month restriction limiting the average work week to 38 hours.

This public choice argument may explain why professions lobby so strongly against retesting those who already have licenses. Even if entry was originally determined by quality, new entrants, because of longer prospective careers, may have more incentive to keep up with the latest innovations. New entrants would therefore tend to do better than some existing practitioners on standardized tests, eliminating some existing licensees.

III. Conclusion

Licensing prevents efficient producers from entering a market by increasing nontransferable rents. The traditional approach thus underestimates the total social cost of licensing by only including the deadweight loss and the direct cost of rent seeking from government cartelization.

REFERENCES

Demsetz, Harold, "Barriers to Entry," *American Economic Review*, March 1982, *72*, 47–57.

Kessel, Reuben A., "Price Discrimination in Medicine," *Journal of Law and Economics*, October 1958, *1*, 20–53.

Leland, Hayne E., "Quacks, Lemons, and Licensing: A Theory of Minimum Quality Standards," *Journal of Political Economy*, December 1979, *87*, 1328–46.

Lott, John R., "The Effect of Nontransferable Property Rights on the Efficiency of Political Markets: Some Evidence," *Journal of Public Economics*, forthcoming 1987.

______, "Brand Names and Barriers to Entry in Political Markets," *Public Choice*, No. 1, 1986, *51*, 87–92.

Texas State Occupational Informational Coordinating Committee, *Directory of Licensed Occupations and Apprenticeship Programs in Texas*, Austin: Texas State Library, September 1983.

On Perfect Rent Dissipation

By JOHN T. WENDERS*

If economists are united on anything, it is the proposition that monopoly prices reduce economic welfare by preventing the realization of the maximum gains from trade in any market. The extent of such distortions to efficiency are often called Harberger costs after Arnold Harberger's 1954 provocative attempt to measure the extent of these losses in the U.S. economy.

More recent analysis has revealed that when monopoly power is achieved via regulation, at least part of the monopoly rents so gained will not be simple transfers from consumers to producers, but will be dissipated by producers' rent-seeking activity. Since such activity employs real resources, there are additional costs to monopolization beyond the Harberger costs as emphasized by Gordon Tullock (1967) and Richard Posner (1975). Indeed, Posner and others have argued that if competition for the monopoly rents is perfect, all of the expected rents from regulation will be converted to welfare losses. While Franklin Fisher's 1985 comment has qualified this conclusion somewhat, the upshot of the debate is that the rent-seeking, or Tullock, costs, may greatly exceed the Harberger costs.[1]

Another recent strand of the analysis concerns the time pattern over which monopoly returns are dissipated by competition to gain and hold the monopoly right. Robert McCormick et al. (1984) emphasize that to the extent such expenditures are sunk, they are forever lost and not recoverable by deregulation. While conceding the point, Martin Cherkes et al. (1986) argue that most rent-seeking expenditures are recurring, not sunk, and therefore large gains from deregulation remain.

The purpose of this essay is to point out that, recurring or sunk, even the largest specification of the Harberger and Tullock costs of regulatory monopolization may fall *far* short of the actual welfare costs. This is because the analysis concentrates on the rent-*seeking* Tullock costs and largely ignores the parallel rent-*defending*[2] Tullock costs. A proper assessment of such rent-defending Tullock costs might more than double the maximum welfare costs of regulation suggested by Posner.

I. Rent-Defending Costs

The general problem with rent-seeking analysis as exposed, for example, by Posner is that it focuses almost entirely on the cost of resources spent by sellers in attaining, and competing with one another for, monopolizing regulation. Those who purchase the output of the industry in question seem to be assumed to sit idly by and await the outcome while sellers scrap over monopoly profits and rents. This is indeed a heroic assumption, and one that was *not* made by Tullock,[3]

*Department of Economics, University of Idaho, Moscow, ID 83843. I am grateful to Richard A. Posner, Gordon Tullock, and two anonymous referees for useful comments on earlier versions of this note. The usual absolution applies.

[1]For an introduction to the growing body of literature in this general area, see the papers by William Corcoran and Gordon Karels (1985), Richard Higgins et al. (1985), and Tullock (1985).

[2]Perhaps a more descriptive term would be "consumers' surplus defending."

[3]This point was recognized by Tullock where, for example, in discussing import tariffs, he says: "One would anticipate that the domestic producers would invest resources in lobbying for the tariff until the marginal return on the last dollar so spent was equal to its likely return producing the transfer. There might also be other interests trying to prevent the transfer and putting resources into influencing the government in the other direction" (1967, p. 228). He also discusses the welfare losses due to resources spent on theft and theft prevention. But this point seems to have been largely lost on the subsequent writers cited. It is also true that Posner, in passing, recognized that there may be rent-defending costs. However, this was not the central point of his paper, which analytically described and measured only the rent-seeking costs. Fisher's comment on Posner

or in the related international trade literature that analyzes trade and tariff restrictions.[4] The point is there are parallel activities and resource expenditures by those who stand to lose from restrictive regulations as they seek to defend against rent-seeking activity and deflect it among themselves.[5] This may lead to a parallel dissipation of consumers' surplus which is additive to the rent-seekers dissipation of producers' surplus.

Because restrictive regulatory practices may appear in a variety of ways, it is difficult to specify generally in the usual supply and demand framework any single measure of the possible extent of rent-defending costs. It depends on the particular restrictive practices proposed. Let us consider a couple of general cases.

First consider an example in which sellers seek to establish a simple binary regulation that would set price at P_M—not necessarily the monopoly profit-maximizing price—in an otherwise competitive market. Consider Figure 1.[6] Here areas H and T represent the standard Harberger efficiency costs and the rent-seeking Tullock costs respectively. Area T represents the maximum benefit that sellers could achieve by setting price P_M, and this, of course, is why it also represents the maximum they will pay in rent-seeking costs to achieve monopolization through regulation.[7]

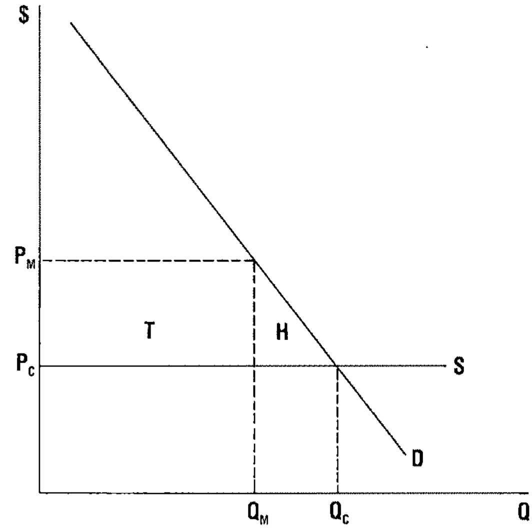


FIGURE 1

On the other hand, the buyers of the product in question would be hurt by an amount equal to their lost consumers' surplus, area $T+H$. Thus, this area represents the maximum buyers would pay in rent-defending activity to avoid the regulation that would set price at P_M.[8] If the regulation under consideration is binary, the admittance of rent-defending activity by buyers means that they have an incentive to use real resources, possibly amounting to a maximum of $T+H$ in Figure 1, to prevent or deflect among themselves the monopolization of the market by the sellers to the degree indicated. Indeed, with perfect rent and surplus dis-

focuses entirely on rent-seeking costs. Finally, while Cherkes et al. mention the possibility that consumers may have a good "incentive to organize politically against the monopoly" (p. 562), they fail to recognize the welfare implications of their observation.

[4]See the survey of endogenous tariff theory by Stephen Magee (1984) and the references cited therein.

[5]A classic example of such activity occurs in public utility rate cases. The utility petitions the commission for a rate increase. Various consumer groups initially present a united front, often with explicit cooperation, in trying to keep the allowed increase in utility revenues as small as possible. However, once the allowed revenue increase is set by the commission, the various consumer groups turn on one another and try to deflect as much of the increase as possible to others. Lions and wolves cooperate in the hunt, but scrap over the kill.

[6]I assume a horizontal supply curve for ease of exposition only. The same general conclusions hold if an upward sloping supply curve, such as that discussed by Harold Demsetz (1984), is employed. In fact, *ceteris paribus*, an upward-sloping supply curve will raise rent-defending costs relative to rent-seeking costs.

[7]If the sellers are able to achieve regulations that allow them to engage in price discrimination, then under Posner's assumptions, even when rent-defending Tullock costs are not admitted, the total welfare costs may exceed $T+H$. If discrimination is perfect, the total welfare costs equal $T+2H$. For simplicity, this point is ignored in the discussion below.

[8]A parallel analysis holds if we look at a situation with an upward-sloping supply curve so that there is the possibility of establishing monopsony power via regulation. Here the amount that sellers would be willing to pay to avoid monopsony would always be greater than the maximum amount buyers would be willing to pay to achieve monopsony.

sipation, the maximum possible welfare costs of successful monopoly would be twice the Tullock and Harberger costs measured by Posner.

Intuitively, this result would seem to double count. If risk-neutral buyers and sellers all believe that the total "prize" at stake is $T+H$, why would the sum of rent-seeking and rent-defending expenditures exceed $T+H$? The answer lies in an analysis similar to the prisoner's dilemma theory (Magee, pp. 47–48). Like the prisoners who both confess, neither buyers nor sellers may refrain from spending the maximum amount they each have at stake. If either voluntarily spends less, they will be taken advantage of by the other side. In addition, there is the incentive for each side to compete *among* themselves to either achieve or avoid the proposed regulation. Thus, both sides may spend up to the amount each has at stake.

Now let us analyze a situation where monopolization is not binary. Suppose sellers are suggesting a form of regulation that would result in a P_M that was profit maximizing, that is, a price that maximized area T, but buyers have the alternative of engaging in rent-defending activities that would hold the regulated price below P_M, say, at P_R, which might lie anywhere between P_M and P_C.

Consider Figure 2. Suppose the sellers initially propose a regulation that would result in price P_M, the full monopoly price. Buyers would be willing to spend an amount up to $a+b$ to water down the proposed regulation so only price P_R came about. For the watered down regulation, sellers would be willing to pay T_0+T_1. Thus the total welfare costs might amount to the entire trapezoid between P_M and P_C, that is, $a+b+T_0+T_1+H$.

This result has a Posner-like characteristic. Posner, and others, have argued competition among sellers for the monopolizing regulation would convert any monopoly rents to rent-seeking Tullock costs so the total welfare costs of regulation can be computed as areas $T+H$ at the *observed* regulated price. When rent-defending Tullock costs are admitted to the analysis, and presuming a parallel kind

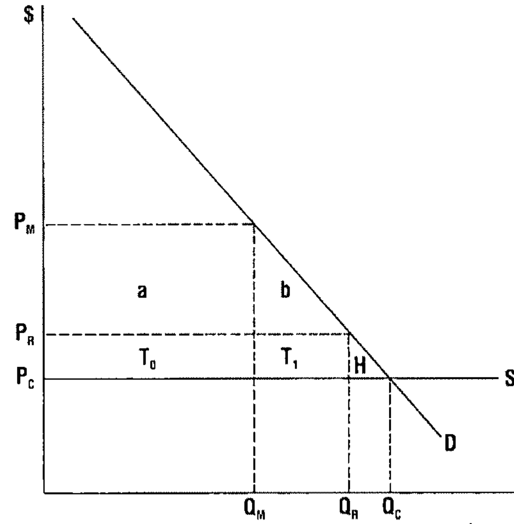


FIGURE 2

of perfect dissipation of the consumers' surplus lost to full monopoly pricing, the total welfare costs amount to a similar, but necessarily larger, Posner-like trapezoid computed at the *unobserved* full monopoly price. The result would be the same as if the regulators always completely ignored the interests of buyers and always gave sellers full monopoly pricing. Under perfect competition among and between buyers and sellers, regulation always results in maximum welfare costs, that is, equal to the rent-seeking Tullock and Harberger costs under full monopoly pricing. And only if sellers succeeded in attaining full monopoly pricing would this result equal the total welfare cost originally proposed by Posner; in all other cases the total welfare loss would be larger due to the admission of rent-defending Tullock costs.

II. Concluding Remarks

1) The above analysis suggests that there is a dissipation of threatened consumers' surplus into rent-defending Tullock costs similar to the dissipation of monopoly rents into rent-seeking Tullock costs. This suggestion needs further analysis along the lines

already explored by Posner, Fisher, and others.

2) Most of the rent-seeking literature presumes that sellers will take the offensive in search of monopoly rents. The above analysis takes this same tack and presumes that buyers are put in the position of defending their consumers' surplus. Yet there is no reason why buyers cannot take the offensive by proposing *monopsonizing* regulation which in turn can be analysed in parallel fashion (James Buchanan, 1975).

3) Any regulation that causes a departure from competitive pricing will benefit the gainers less than it costs the losers. Indeed, this is what causes Harberger costs. If buyers and sellers stakes were equally potent, dollar for dollar, in purchasing regulatory influence, efforts to achieve monopoly through regulation would never be successful because the losers could always outspend the gainers. The general task, then, of a positive economic theory of regulation is to explain how the various degrees of asymmetry in regulatory influence between buyers and sellers result in balances at the margin that produce the observed degree of regulation.

4) More generally, in an environment where regulation is constitutionally admitted, the entire gains from trade in any market, as depicted by the triangle between the demand and supply curves up to their point of intersection, is vulnerable to various kinds of rent-seeking, rent-defending, and political extortion.[9] All of these can greatly increase welfare costs above the simple Harberger variety.

At one time, the cost of monopoly was neatly settled and made its way into even elementary texts. By now, it should be clear that these matters are far from settled. A new, stable, orthodoxy in the area is still a long way off.

[9]Fred McChesney (1987) emphasizes that politicians are not merely passive brokers in the regulatory process, but are also independent actors making their own demands on the available gains from trade by threatening regulation or taxation.

REFERENCES

Buchanan, James M., "Consumerism and Public Utility Regulation," in Charles F. Phillips, Jr., ed., *Telecommunications, Regulation, and Public Choice*, Lexington, Washington and Lee University, 1975, 1–22.

Cherkes, Martin, Friedman, Joseph and Spivak, Avia, "The Disinterest in Deregulation: Comment," *American Economic Review*, June 1986, *76*, 559–63.

Corcoran, William J. and Karels, Gordon V., "Rent-Seeking Behavior in the Long-Run," *Public Choice*, No. 3, *46*, 227–46.

Demsetz, Harold, "Purchasing Monopoly," in David C. Collander, ed., *Neoclassical Political Economy*, Boston: Ballinger, 1984, 101–14.

Fisher, Franklin M., "The Social Costs of Monopoly and Regulation: Posner Reconsidered," *Journal of Political Economy*, April 1985, *93*, 410–16.

Harberger, Arnold C., "Monopoly and Resource Allocation," *American Economic Review Proceedings*, May 1954, *44*, 77–87.

Higgins, Richard S., Shughart, William F. II and Tollison, Robert D., "Free Entry and Efficient Rent-Seeking," *Public Choice*, No. 3, 1985, *46*, 247–58.

McChesney, Fred S., "Rent Extraction and Rent Creation in the Economic Theory of Regulation," *Journal of Legal Studies*, January 1987, *16*.

McCormick, Robert E., Shughart, William F. II and Tollison, Robert D., "The Disinterest in Deregulation," *American Economic Review*, December 1984, *74*, 1075–79.

Magee, Stephen C., "Endogenous Tariff Theory: A Survey," in David C. Collander, ed., *Neoclassical Political Economy*, Boston: Ballinger, 1984, 41–51.

Posner, Richard A., "The Social Costs of Monopoly and Regulation," *Journal of Political Economy*, August 1975, *83*, 807–27.

Tullock, Gordon, "The Welfare Costs of Tariffs, Monopolies and Theft," *Western Economic Journal*, June 1967, *5*, 224–32.

______, "Back to the Bog," *Public Choice*, No. 3, 1985, *46*, 259–63.

A Note on Bilateral Monopoly and Formula Price Contracts

By ROGER D. BLAIR AND DAVID L. KASERMAN*

A bilateral monopoly exists when an upstream monopolist sells its output to a downstream monopsonist. In these circumstances, vertical integration is profitable because it eliminates the need for repeated and protracted (and, therefore, costly) negotiations over the price and quantity of the intermediate product.[1] As a result, vertical integration increases the combined profits of the two monopolists (George Stigler, 1966, p. 208).

For a variety of reasons, however, ownership integration may not be an attractive alternative.[2] As a result, bilateral monopolists may seek contractual arrangements that are economically equivalent to vertical integration.[3] In this paper, we derive a contractual agreement that achieves results that are very close to ownership integration in the bilateral monopoly situation. This agreement specifies a formula that determines the intermediate product price as a function of the final output price and the average costs of production at both the upstream and the downstream stages. It has several desirable features: 1) it leads each firm to pursue independent profit-maximizing policies that will result in maximum joint profits; 2) it allocates a specified share of these maximized joint profits to each party to the contract; and 3) it automatically adapts to changes in production costs and final output demand.

I. Formula Price Contracts

To derive the contractual alternative to ownership integration in the bilateral monopoly situation, we use the following notation and assumptions:

Q = output of the final product; X = quantity of the intermediate product; $Q = X$, assume a fixed input/output ratio$=1$; $C_X(X)$ = average cost of producing input X; $C_T(Q)$ = average cost of transforming one unit of X into one unit of Q; $P(Q)$ = final output demand; and $p_X(X)$ = price of input X.

Now, assume that there exists a bilateral monopoly in the market for X.[4] That is, assume that the sale of X is monopolized

*Department of Economics, College of Business Administration, University of Florida, Gainesville, FL 32611, and College of Business, Auburn University, Auburn, AL 36849–3501, respectively. We are grateful to the Public Policy Research Center at the University of Florida for financial support. Jay Creswell provided constructive comments during the formative stages of this paper. The usual caveat applies.

[1] A. L. Bowley (1928) recognized that bilateral monopolists will agree to exchange the joint profit-maximizing quantity of the intermediate product. Otherwise, the firms will not be on the contract curve. Given this quantity, the firms must then negotiate the price of the intermediate good. This price will determine how the maximized joint profits are divided between the two stages of production. For an interesting survey of the early literature dealing with bilateral monopoly, see Fritz Machlup and Martha Taber (1960). As Oliver Williamson (1971, 1974) explains, each firm will expend resources on the price negotiation process up to the point at which the marginal cost of further negotiation equals the marginal revenue that such negotiation yields. Such expenditures represent a drain on the total profits available. Thus, firms may vertically integrate to avoid this negotiation.

[2] Vertical integration may involve managerial diseconomies (Williamson, 1973), increased capital costs (Williamson, 1974), and substantial costs of negotiating the price of the acquired firm.

[3] On contractual alternatives generally, see our 1983 study. Several authors have recognized that the vertical relationships we observe in practice are not always easily categorized according to a binary market versus nonmarket taxonomy. See Benjamin Klein et al. (1978), K. J. Blois (1972), and Victor Goldberg (1979).

[4] The analysis we present here applies equally well to the successive monopoly model in which both the X market and the Q market are monopolized with no monopsony power in the purchase of X. The model, however, does appear to require that fixed proportions obtain in the production of X.

and the purchase of X is monopsonized. In the absence of vertical integration, the upstream monopolist's profit function will be given by

$$(1)\qquad \Pi_U = p_X(X)X - C_X(X)X,$$

and the downstream monopsonist's profit function will be

$$(2)\qquad \Pi_D = P(Q)Q - p_X(X)X - C_T(Q)Q.$$

If these two firms were to vertically integrate, the profit function of the combined operation would be

$$(3)\qquad \Pi_I = P(Q)Q - C_X(X)X - C_T(Q)Q.$$

Suppose, however, that these two firms do not wish to vertically integrate. Instead, they would like to sign a long-term agreement that will generate combined profits equal to Π_I^* (the maximized value of equation (3)) and automatically assign shares of these profits equal to α and $1-\alpha$ to the upstream firm and downstream firm, respectively, where $0 \le \alpha \le 1$. Obviously, such an assignment of profit shares must occur through the price of the intermediate product that is determined by the contract.

Setting $\Pi_U = \alpha\Pi_I$ and solving for p_X, we obtain

$$(4)\qquad p_X = \alpha(P - C_T) + (1-\alpha)C_X.$$

Thus, if the upstream firm can ensure that the price of the intermediate product is determined by final output price and production costs at both stages through the formula given in equation (4), it will be assured of receiving α of the profits available to a vertically integrated monopolist.[5] Moreover, substituting equation (4) into equation (2), we find

$$(5)\qquad \Pi_D = PQ - [\alpha(P - C_T) + (1-\alpha)C_X] \times X - C_T Q = (1-\alpha)\Pi_I.$$

That is, the formula price contract described in equation (4) also automatically assigns $1-\alpha$ of the integrated profits to the downstream firm.

Therefore, under the terms of the formula price contract of equation (4), $\Pi_U = \alpha\Pi_I$ and $\Pi_D = (1-\alpha)\Pi_I$. As a result, independent profit maximization by the two nonintegrated firms will lead to combined profits of Π_I^*.[6] Thus, the formula price contract provides a viable alternative to repeated negotiations and to ownership integration in the bilateral monopoly situation.

II. Performance Characteristics

The formula price contract appears to provide a reasonably close substitute for vertical integration. It does this by providing a degree of flexibility that is equal to or greater than that obtained with a series of short-term contracts while still defining long-term supply and demand obligations that do not require periodic renegotiation. Such contracts exhibit three desirable performance characteristics.

First, they facilitate the negotiation process by focusing attention on a single parameter, α. With each firm automatically driven to produce the joint profit-maximizing level of output, there is no need to specify the price and output of the intermediate product. The parties to the contract need only settle on mutually agreeable shares of the resulting maximized profit.

Second, the formula price contract further facilitates negotiation by economizing on the information needed for contract specification. Again, because of the incentive struc-

[5]Equation (4) may be rewritten as

$$p_X = C_X + \alpha(P - C_T - C_X).$$

Thus, the intermediate product price equals the average cost of producing the input plus α times the (optimal) integrated monopoly markup over cost at the downstream stage.

[6]Proof of this statement is straightforward and may be obtained from the authors upon request.

ture that automatically leads firms to produce the joint profit-maximizing output, the parties to the contract do not require information on final product demand in order to specify the terms of the contract. That is, the contract can be negotiated without specific knowledge of $P(Q)$.

Third, once the formula price contract is in effect, any changes in final output demand or in production costs at either stage will be reflected appropriately in the new optimal prices and outputs fostered by the contract. That is, the contract will automatically accommodate changes in $P(Q)$, $C_T(Q)$, or $C_X(X)$, encouraging the parties to the contract to adjust to the new Π_I^* solution. Moreover, the firms' shares in this new Π_I^* will remain at the agreed upon values of α and $1-\alpha$. No renegotiation is required as a result of dynamic changes in final product demand or in costs.[7] In this respect, the formula price contract appears to operate in a fashion that is identical to the ownership integration alternative.

III. Strategic Misrepresentation Risk

We can see from equation (4) that the intermediate good price is a function of the unit cost of production at both stages. This raises the possibility that the formula price contract offers the parties postnegotiation incentives to behave opportunistically by misrepresenting $C_X(X)$ or $C_T(Q)$. To the extent that incentives for strategic misrepresentation exist, some sort of audit or policing mechanism may be required, thereby raising the transaction costs of employing these contracts.

Unfortunately, there is an incentive for some strategic misrepresentation of costs after the formula price contract is in effect. From equation (4), if the upstream firm's true average costs are C_X but the firm claims that they are $C_X+\delta$ where $\delta \geq 0$, the intermediate product price determined by the contract will be

$$(6) \quad p_X = \alpha(P - C_T) + (1-\alpha)(C_X+\delta) \ .$$

Note that final output price, P, will be affected by the markup on costs, δ. Substituting (6) into equation (1) and recalling that $X = Q$, the upstream firm's profits with misrepresentation are

$$(7) \quad \begin{aligned} \Pi_U &= [\alpha(P - C_T) + (1-\alpha)(C_X+\delta)] \\ &\quad \times X - C_X X \\ &= \alpha[PQ - C_T Q - (C_X+\delta)X] + \delta X \\ &= \alpha \Pi_I(\delta) + \delta Q(\delta). \end{aligned}$$

In other words, the upstream firm's profits are its share of the *new* joint profit, $\Pi_I(\delta)$, which is a function of the level of misrepresentation and is necessarily less than Π_I^*, plus the markup on cost times the *new* output, which is also a function of δ.

Differentiating Π_U with respect to δ and evaluating the sign at $\delta = 0$, indicates that $\partial \Pi_U / \partial \delta > 0$ at $\delta = 0$. That is, the upstream firm has an incentive to overstate its costs for all $\alpha < 1$. An analogous incentive to misrepresent costs exists for the downstream firm as well.[8] An interesting issue, then, involves the contractual and noncontractual responses to such incentives for strategic behavior. A thorough investigation of this issue is beyond the scope of this paper. At this point, we merely indicate that, since both parties to the contract may engage in misrepresentation, there will be a clear incentive to incorporate some sort of policing mechanism (such as postnegotiation audits) in the contractual arrangement. Alternatively, the relational aspects of this sort of long-term contract together with the potential for retaliation may be sufficient to

[7]This is not to say that renegotiation of the profit shares themselves will never be necessary. But fluctuations in cost and demand conditions, which may occur frequently, do not require renegotiation of α.

[8]The optimal level of misrepresentation for the upstream firm is

$$\delta^* = -[\alpha \partial \Pi_I / \partial \delta + Q] / (\partial Q / \partial \delta) > 0.$$

This derivation, along with proofs that $\partial \Pi_U / \partial \delta$, $\partial \Pi_D / \partial \delta > 0$ at $\delta = 0$, are available from the authors upon request.

mitigate the incentives to behave opportunistically.

IV. Conclusion

In this paper, we have demonstrated a fundamental equivalence between ownership integration and formula price contracts in the bilateral monopoly model. Thus, there is an attractive contractual alternative to vertical ownership integration in this market setting. Consequently, the choice between the contractual and ownership alternatives should turn on a comparison of the relative costs associated with each, which we have abstracted from here. Formula price contracts require negotiation on profit shares and some sort of postnegotiation audits of the production costs at both stages. Ownership integration, in contrast, requires negotiation on the terms of the merger or acquisition and may entail certain managerial or pecuniary (capital cost) diseconomies. Clearly, neither alternative will dominate in all instances.

REFERENCES

Blair, Roger D. and Kaserman, David L., *Law and Economics of Vertical Integration and Control*, New York: Academic Press, 1983.

Blois, K. J., "Vertical Quasi-Integration," *Journal of Industrial Economics*, July 1972, *20*, 253–72.

Bowley, A. L., "Bilateral Monopoly," *Economic Journal*, December 1928, *28*, 651–59.

Goldberg, Victor, "The Law and Economics of Vertical Restrictions: A Relational Perspective," *Texas Law Review*, December 1979, *58*, 91–129.

Klein, Benjamin, Crawford, Robert G. and Alchian, Armen A., "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

Machlup, Fritz and Taber, Martha, "Bilateral Monopoly, Successive Monopoly, and Vertical Integration," *Economica*, May 1960, *27*, 101–19.

Stigler, George J., *The Theory of Price*, 3rd ed., New York: Macmillan, 1966.

Williamson, Oliver E., "The Vertical Integration of Production: Market Failure Considerations," *American Economic Review Proceedings*, May 1971, *61*, 112–23.

______, "Markets and Hierarchies: Some Elementary Considerations," *American Economic Review Proceedings*, May 1973, *63*, 316–25.

______, "The Economics of Antitrust: Transaction Cost Considerations," *University of Pennsylvania Law Review*, May 1974, *122*, 1439–96.

Seasonality, Aggregation and the Testing of the Production Smoothing Hypothesis

By MOHEB A. GHALI*

One of the leading hypotheses concerning the dynamics of production over time is the production smoothing hypothesis. Given a planning horizon which spans a number of production periods, the firm need not produce in each period an amount equal to expected sales. Rather, resorting to inventory accumulation and liquidation, the firm may follow a production plan temporally smoother than the path of demand. If firms faced with convex cost functions chose to smooth the rate of output in order to minimize costs, one would expect to observe that the rate of output would vary less than the rate of sales, with variations in inventory stocks absorbing some of the fluctuations in sales. Recently, work on the testing of the production smoothing hypothesis has cast doubt on its empirical validity. The evidence presented by Alan Blinder seems to indicate that the variance of production exceeds that of sales in seven out of eight two-digit retail industries (1981) and in eighteen out of twenty two-digit manufacturing industries (1983 and 1986).

The purpose of this paper, then, is to examine the validity of such tests when seasonally adjusted aggregated data are used. The evidence presented show that the relative size of the variances of the seasonally adjusted production and sales does not provide valid tests of the production smoothing hypothesis. In addition, aggregating over firms where the seasonal patterns differ may also distort tests of production smoothing. Blinder realized that the use of seasonally adjusted data may not provide an adequate test of the hypothesis, stating "Had they been available, I would have preferred to use data that were not seasonally adjusted since the production smoothing model presumably applies to seasonal fluctuations in sales. However, such data are not available" (1983, fn. 19).

In this paper I focus on the cement industry because the unadjusted disaggregated data are available for the direct testing of the conjecture that aggregate seasonally adjusted data mask production smoothing phenomenon. Aggregate monthly data on five other industries will also be examined.

I. Output, Inventories, and Sales

The effects of seasonal adjustment and aggregation on the testing of the production smoothing hypothesis can be shown by examining the descriptive statistics calculated from the unadjusted and the adjusted data. This I shall do and report the results below. The distortions induced by data adjustment are even more clear when examined in the context of firms' behavior. Seasonally adjusted data may lead to the wrong inference regarding the firms' behavior. To illustrate this point, I use two sinple models, each representing the short-run behavior of output and inventories. My objective is to examine the effects of the seasonal adjustment under different models rather than to introduce a new model or to provide a comparative evaluation of the models. The simplicity of the models chosen and of the assumptions used should be viewed with this objective in mind.

For both models, I assume that the firm has a planning horizon of a specific length and which consists of N production periods.[1]

*Professor of Economics, University of Hawaii, Honolulu, Hawaii 96822. I am grateful to two referees for valuable comments.

[1] The start of the planning horizon is taken as: "... the period after the peak demand period when the demand rate first falls below the average for the cycle, the period where inventories characteristically reach minimum" (T. Magee and D. M. Boodman, 1967, pp. 166–67). See

The cost functions are assumed quadratic and are defined per production period. Should the firm decide to increase the stock of inventories held at the end of the planning horizon I_N, the desired increase $(I_N - I_0)$ is viewed as additional demand for output in period N.

The first of the simple production smoothing models is that of my earlier papers (1974, 1981). As a result of minimizing costs over the horizon output in production period i is expressed as a weighted average of expected sales in that period, S_i, and the expected average sales over the planning horizon, $\bar{S}^{*e}$:

$$(1) \qquad P_i = (1-\alpha)S_i + \alpha\bar{S}^{*e}.$$

Actual sales are written as the sum of expected sales and a random disturbance, e_i, which has zero expected value. The firm is assumed not to revise its expectations during the horizon. The production plan can then be written as

$$(1') \qquad P_i = (1-\alpha)S_i + \alpha\bar{S} - e_i',$$

where $\bar{S}$ is the average sales over the horizon measured as: $1/N[(\Sigma_{i=1}^{N}S_i)+(I_N - I_0)]$ and the disturbance $e' = (1-\alpha)e_i + \alpha\bar{e}$, where $\bar{e} = (1/N)\Sigma_{i=1}^{N}e_i$.

The coefficient α is a measure of the degree of production smoothing: when $\alpha = 1$, output will proceed at a perfectly smooth pace, and when $\alpha = 0$, planned output in each period will equal expected sales for that period. This degree of smoothing depends on the cost conditions facing the firm but will be bounded by $0 \leq \alpha \leq 1$ (see my 1982 paper).

The second model I use to illustrate the effects of seasonal adjustment is that of equation (1.3) of Blinder's paper (1983). Inventory investment is written as the sum of two components: a fraction β_1 of the desired change in stock, and a fraction $(1-\beta_2)$ of the unanticipated sales. In terms of the notation, I have used thus far:

$$(2) \qquad \Delta I_i = \beta_1(I_i^* - I_{i-1}) - (1-\beta_2)(S_i - S_i^e).$$

The coefficient $(1-\beta_2)$ has been interpreted as the degree of smoothing: as $(1-\beta_2)$ approaches unity (i.e., as β_2 approaches zero), all unanticipated sales will be met through variations in inventories rather than output. It would be more appropriate to call it "the degree of buffering" to distinguish it from the planned variations in inventories in response to anticipated variations in sales.[2] To fit this model to observed data, assume that the desired stock is determined by a flexible accelerator relationship (see Michael Lovell, 1964; George Hay, 1970; and Owen Irvine, 1981):

$$(3) \qquad I_i^* = \gamma_0 + \gamma_1 S_i^e.$$

Output can then be written as a function of sales, expected sales, and lagged inventory stock:

$$(4) \qquad P_i = \gamma_0\beta_1 + \beta_2 S_i + (1+\gamma_1\beta_1 - \beta_2)S_i^e - \beta_1 I_{i-1}.$$

A distributed lag function with geometrically declining weights, λ, is used for expected sales (John Muth, 1985). With a Koyck transformation, the buffer stock model is written as

$$(5) \qquad P_i = \delta_0 + \delta_1 S_i - \delta_2 S_{i-1} - \delta_3 I_{i-1} + \delta_4 I_{i-2} + \delta_5 P_{i-1},$$

where $\delta_1 = 1 + \gamma_1\beta_1$, $\delta_2 = \lambda\beta_2$, $\delta_3 = \beta_1$, $\delta_4 = \lambda\beta_1$ and $\delta_5 = \lambda$.

II. The Effects of Seasonal Adjustment and Aggregation

In Table 1, I report the statistics for the producers of the portland cement industry, using the data used in my earlier studies

also Morton Klein (1961), Franco Modigliani and Franz Hohn (1955) and Modgliani and Owen Sauerlender (1955). For portland cement, the planning horizon was assumed to begin at the start of November.

[2]This distinction is consistent with the definitions of "buffering" and "smoothing" in Richard Ashley and Daniel Orr (1985).

TABLE 1—MONTHLY DATA ON PORTLAND CEMENT AND FIVE OTHER INDUSTRIES[a]

	Unadjusted		Detrended		Seasonally Adjusted Detrended	
Industry	$V(P)/V(S)$	$\text{Cor}(\Delta I, S)$	$V(P)/V(S)$	$\text{Cor}(\Delta I, S)$	$V(P)/V(S)$	$\text{Cor}(\Delta I, S)$
Portland Cement Prod. District:						
1	.42	−.780	.41	−.779	1.08	−.295
2	.34	−.853	.32	−.869	.84	−.325
3	.48	−.766	.42	−.808	1.18	−.209
4	.28	−.887	.28	−.840	.87	−.475
5	.62	−.687	.59	−.716	1.03	−.304
6	.15	−.932	.15	−.933	.55	−.622
7	.50	−.806	.42	−.836	1.23	−.198
8	.83	−.536	.75	−.566	.90	−.416
9	.63	−.675	.59	−.678	1.00	−.428
10	1.03	−.156	1.06	−.371	1.52	−.146
11	.22	−.878	.16	−.894	.43	−.564
12	.31	−.886	.27	−.905	.59	−.622
13	.39	−.675	.40	−.680	.80	−.387
14	.58	−.688	.50	−.742	.94	−.364
15	.89	−.325	.84	−.415	.91	−.358
16	.74	−.546	.62	−.658	.88	−.407
17	.83	−.510	.76	−.639	1.03	−.501
18	.64	−.659	.64	−.655	.83	−.483
19	1.00	−.298	.67	−.368	1.00	−.389
Aggregate	.47	-.830	.38	−.879	1.00	−.279
Asphalt	.50	−.869	.45	−.894	.92	−.303
Oil Burners	.79	−.502	.70	−.597	1.17	−.050
Glass Containers	.60	−.642	.42	−.771	.32	−.826
Printing Paper	1.11	.075	1.21	.094	1.22	.054
Beer	1.04	−.112	1.05	−.106	1.17	.104

[a] The production districts are 1) E. Pennsylvania, Maryland; 2) New York, Maine; 3) Ohio; 4) W. Pennsylvania, West Virginia; 5) Michigan; 6) Illinois; 7) Indiana, Kentucky, Wisconsin; 8) Alabama; 9) Tennessee; 10) Virginia, South Carolina, Georgia, Florida, Louisiana, Mississippi; 11) Iowa; 12) E. Missouri, Minnesota, South Dakota; 13) Kansas; 14) W. Missouri, Nebraska, Oklahoma, Arkansas; 15) Texas; 16) Colorado, Arizona, Utah, Wyoming, Montana, Idaho; 17) California; 18) Oregon, Washington; 19) Puerto Rico.

(1981; 1982). The industry is divided into production districts, nineteen of which have continuous monthly data on output and sales (in physical quantities). I use 120 observations for each of the districts covering the period 1950–60.

In all but one case (the tenth district, covering Virginia, South Carolina, Georgia, Florida, Louisiana, and Mississippi) the variance of production is less than that of sales and in all cases the correlation between sales and inventory change is negative. These results are exactly what one expects under the production smoothing hypothesis.

To separate the effect of detrending from that of seasonal adjustment, I first detrended the data. The results, reported in Table 1 show the same pattern as those of the unadjusted data. The variance of production is less than the variance of sales for all but district 10, and all the coefficients of correlation between sales and inventory change are negative.

The data were then detrended and seasonally adjusted by regressing each of the variables on eleven monthly dummy variables and a trend variable (see Lovell, 1963). The ratio of the variances and the covariances of the adjusted variables are also reported in the table.

It will be noted that while the coefficients of correlation between sales and inventory change remain negative, they have decreased. More important, in eight cases out of the nineteen, the variance of production was equal to or greater than the variance of

sales. The ratio of the variance of production to the variance of sales exceeded 80 percent in fifteen cases. If one were to judge the production smoothing hypothesis on the basis of seasonally adjusted data, one would have to conclude that there seems to be no strong evidence in support of the hypothesis. This is clearly an incorrect conclusion, given the results of the unadjusted data.

When the data are aggregated into a single time-series, we obtain the results reported in the aggregate row of the table. The conclusions regarding seasonal adjustment and the evidence on smoothing are the same as for diseggregated data.

In my earlier study (1982), monthly data on five industries other than portland cement were analyzed. The production smoothing model was found to provide an explanation of monthly output behavior in three industries: asphalt, oil burners, and glass containers (in addition to cement). The model failed, however, to explain the monthly variations in output in the printing paper and beer industries.[3] I used the same data, pooling the 120 observations for each industry and using monthly dummy variables and a trend variable to detrend and seasonally adjust the data as I have done for the cement data.

Table 1 reports the ratio of the variance of output to the variance of sales, and the covariance of inventory investment and sales for both the original and the seasonally adjusted, detrended data.

The variance of output is less than that of sales for the first three industries for which my earlier 1982 study had found significant production smoothing. For the printing paper and the beer industries, the variations in production exceed those in sales, though the ratios are close to unity. For the first three industries the coefficients of correlation between sales and inventory investment are negative and high. For beer, the correlation is negative, but small. For printing paper, the correlation is positive but small.

When seasonally adjusted detrended data are used, the ratio of the variance of output to that of sales increases for all but the glass containers industry. Given the closeness of the ratio to unity for the asphalt industry, one would have to conclude that production smoothing would have been masked if the study had relied upon seasonally adjusted data.

The data are next used to estimate the models directly, and to examine the effects of seasonal adjustment and aggregation on inference regarding firms' behavior. The results obtained for equation (1′) using the seasonally unadjusted data were reported in my earlier paper (1982, Table II) and are reproduced here in the left-hand side of Table 2 for comparison with those reported in the right-hand side using the seasonally adjusted detrended data.[4] Though the model imposes some restrictions on the parameters, I estimated the equations with no restrictions and use the degree to which the results satisfy the theoretical restrictions as a measure of the model's performance. Specifically, the intercept should be zero and the coefficients of sales and average sales should add up to unity.

The results reported in Table 2 give production smoothing strong support. All but two of the estimated degree of production smoothing α are significantly different from zero at the 1 percent level, and the two prior constraints on the intercept and the sum of the coefficients are satisfied in all cases.

Turning now to the results obtained by fitting the model to seasonally adjusted data, any judgment concerning the validity of the production smoothing model must be reversed. In fifteen of nineteen regressions, the degree of production smoothing does not differ from zero at the 1 percent level, and in all cases current production is significantly influenced by current sales. While fifteen out of nineteen cases can be viewed as sufficient grounds for dismissing production smoothing as an empirically useful hypothesis, I

[3] The start of the planning horizons for the five industries are: asphalt: November; beer: January; glass containers: February; oil burners: December; printing paper: February.

[4] In reporting the results of these regressions, are the results of significance testing are footnoted. Full results with the estimated standard errors of coefficients, analysis of variance results, and test statistics for autocorrelation are available from the author.

TABLE 2—THE PRODUCTION SMOOTHING MODEL

$P_i = c + (1-\alpha)S_i + \alpha\bar{S} + e_i'$

	Unadjusted Data					Seasonally Adjusted, Detrended Data				
Industry	c	$(1-\alpha)$	(α)	R^2	t for H_0	c	$(1-\alpha)$	(α)	R^2	t for H_0
Portland Cement Prod. Districts:										
1	.61	.50[a]	.33[a]	.64	−1.40	−.42	.80[a]	.14	.63	−.60
2	−.14	.49[a]	.63[a]	.82	.99	−.22	.80[a]	.15	.81	−.66
3	−.04	.56[a]	.49[a]	.82	.55	−.02	.87[a]	.01	.62	−1.34
4	−.10	.41[a]	.71[a]	.75	.99	−.49[a]	.53[a]	.51*	.57	.36
5	−.04	.72[a]	.31[a]	.91	.39	−.04	.78[a]	.29	.61	−2.40
6	−.10	.31[a]	.86	.65	.47	−.10	.40[a]	.13	.32	−1.75
7	−.23	.59[a]	.61[a]	.85	2.04	0	.93[a]	0	.70	−.89
8	−.17	.67[a]	.54[a]	.79	2.39	−.25[a]	.67[a]	.26[a]	.59	−.80
9	−.13	.60[a]	.64[a]	.68	1.40	−.26	.65[a]	.42[a]	.54	.54
10	−.01	.72[a]	.30[a]	.89	.61	−.03	.79[a]	.02	.42	−2.09
11	−.08	.33[a]	.79[a]	.76	1.29	−.05	.42[a]	−.06	.42	−7.01[a]
12	−.09	.47[a]	.63[a]	.84	1.00	−.16	.51[a]	.15	.49	−3.87[a]
13	−.09	.49[a]	.62[a]	.75	.97	−.44[a]	.49[a]	.59[a]	.59	.78
14	−.03	.62[a]	.43[a]	.85	.70	−.07	.76[a]	.08	.63	−2.32
15	−.02	.71[a]	.31[a]	.85	.37	−.19[a]	.74[a]	.10	.68	−2.68[a]
16	−.01	.72[a]	.30[a]	.92	.48	−.02	.66[a]	.02	.48	−4.64[a]
17	−.02	.70[a]	.29[a]	.84	−.27	−.11	.71[a]	.04	.50	−3.26[a]
18	0	.73[a]	.27	.87	.00	−.28	.52[a]	.46[a]	.58	−.22
19	0	.72[a]	.28[a]	.87	.03	−.03	.80[a]	.07	.75	−2.65[a]
Aggregate	−.97	.56[a]	.49[a]	.89	.90	−.50	.90[a]	.02	.72	−1.29
Asphalt	.23	.67[a]	.30[a]	.95	−.53	−.10	.88[a]	.10	.84	−1.91
Oil Burners	−1.00	.76[a]	.26[a]	.93	.44	−1.45	.97[a]	.03	.81	.08
Glass Containers	.42	.49[a]	.49[a]	.82	−.52	−.30	.28[a]	.02	.25	−12.12
Printing Paper Paper	0	1.04[a]	−.03	.93	.13	.01	1.02[a]	.0	.86	.43
Beer	.13	.96[a]	.02	.89	−.10	.13	1.03[a]	−.02	.92	.18

Note: The "t for H_0" column is the t value for the hypothesis $(1-\alpha)+\alpha=1$.

[a]Denotes significant at the 1 percent level.

would have had no hesitation at all rejecting it had I estimated the model for the aggregated data only. With a sales coefficient of .89, which does not differ significantly from unity, I would have concluded that producers attempt to produce an output equal to sales every production period, and that the production smoothing hypothesis, though theoretically attractive, is devoid of empirical significance.

The same conclusion is obtained when the results of direct estimation of the degree of production smoothing for the five industries are examined. When the unadjusted data are used, the degree of production smoothing significantly differed from zero, ranging from .26 to .49 for the first three industries. The theoretical restrictions on the coefficients are satisfied for these industries. The degree of smoothing is zero for printing paper and beer, where output in each month is equal to sales (the coefficient of sales does not significantly differ from unity). Use of seasonally adjusted detrended data renders all estimates of the degree of smoothing insignificant.

The attempt to estimate the buffer stock model, equation (5) was not successful. The high collinearity between the variables and their lagged values resulted in unrealistic speeds of adjustment coefficients, eight of which differed significantly from zero and ranged from .48 to 5.09. Similarly, the coefficient of lagged sales, measuring $\lambda\beta_2$ ranged from −.32 to −4.9. When the seasonally adjusted data were used, the explanatory power of the model was reduced and only in one of the twenty cases did the estimated speed of adjustment or the coefficient $\lambda\beta_2$

differ from zero. These results did not permit me to draw inference regarding β_2. (The tabulated results are available from the author.)

IV. Conclusion

Cost minimization by firms facing convex cost functions will result in production plans in which variations in the rate of output are small relative to those in expected sales. This is made possible by the accumulation and the subsequent liquidation of inventories. The degree to which firms will smooth the rate of output depends on the relative costs of changing the rate of output as opposed to changing stock level. This logical implication of cost function convexity was elegantly derived for quadratic functions by Charles Holt et al. (1960). In addition to convexity of cost function, the ability of the firm to forecast the variations in sales over the planning horizon is crucial to its ability to follow a production smoothing plan. For a planning horizon of a year, knowledge of the seasonal pattern of demand (and the degree of temporal stability of that pattern) determine the firms success in minimizing cost through production smoothing. If we choose to extract seasonality from the data, the ratio of the variance of production to that of sales will not be an indicator of production smoothing, nor can inference be drawn from such a ratio concerning the convexity of the cost functions or the firm's dynamic behavior.

REFERENCES

Ashley, Richard and Orr, Daniel, "Further Results on Inventories and Price Stickiness," *American Economic Review*, December 1985, *75*, 964–75.

Blinder, Alan, "Retail Inventory Behavior and Business Fluctuations," *Brookings Papers on Economic Activity*, 2:1981, 443–505.

______, "Can the Production Smoothing Model of Inventory Behavior Be Saved?," mimeo., *Econometric Society Meetings*, December 1983.

______, "Can the Production Smoothing Model of Inventory Behavior Be Saved?," *Quarterly Journal of Economics*, August 1986, *101*, 431–53.

Ghali, Moheb, "Inventories, Production Smoothing and the Accelerator: Some Empirical Evidence," *Quarterly Journal of Economics*, February 1974, *88*, 149–57.

______, "Production Smoothing and Inventory Behavior: A Simple Model," in A. Chikan, ed., *Economics and Management of Inventories*, Amsterdam: Elsevier, 1981.

______, "Inventories and Short-run Output Stabilization," *Southern Economic Journal*, January 1982, *48*, 614–26.

Hay, George, "Adjustment Costs and the Flexible Accelerator," *Quarterly Journal of Economics*, February 1970, *84*, 140–43.

Holt, Charles et al., *Planning Production Inventories and Workforce*, Englewood Cliffs: Prentice Hall, 1960.

Irvine, Owen, "Retail Inventory Investment and the Cost of Capital," *American Economic Review*, September 1981, *71*, 633–48.

Klein, Morton, "On Production Smoothing," *Management Science*, April 1961, *7*, 286–93.

Lovell, Michael, "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of the American Statistical Association*, December 1963, *58*, 993–1010.

______, "Determinants of Inventory Investment," in *Models of Income Determination*, NBER *Studies in Income and Wealth*, Vol. 28, University Microfilms, 1964, 177–244.

Magee, T. and Boodman, D. M., *Production Planning and Inventory Control*, 2nd ed., New York: McGraw-Hill, 1967.

Modigliani, Franco and Hohn, Franz, "Production Planning Over Time and The Nature of Expectations and Planning Horizon," *Econometrica*, January 1955, *23*, 46–66.

______ **and Sauerlender, Owen,** "Economic Expetations and Plans of Firms in Relation to Short-Term Forecasting," in *Short-Term Economic Forecasting*, NBER *Studies in Income and Wealth*, Vol. 17, University Microfilms, 1955, 261–351.

Muth, John, "Short-Run Forecasts of Business Activity," paper presented at the AEA meetings, New York, December 1985.

The Validity of ROI as a Measure of Business Performance

By Robert Jacobson*

Business performance at the corporate level is widely assessed by the return, through changes in the price of the stock and dividends, to shareholders. However, corporate level performance analysis, whether for strategic or public policy purposes, is inappropriate in many instances because of the heterogeneity of the corporation's operations. The strategic business units (SBUs) comprising a corporation are often very diverse. Profitability analysis grouping these entities together yields few of the insights needed for strategic or public policy decision making.

Because of this heterogeneity, profitability analysis is often advocated and undertaken at the SBU level. The absence of an equity market at the SBU level means that a measuring instrument other than that based on stock price must be used to assess profitability. Due to the lack of alternatives, the performance of an SBU is almost always based on accounting data. In particular, the accounting measure, return on investment (ROI), is widely regarded as the most useful measure and ultimate "bottom line" test of business performance (James Reese and William Cool, 1978). It is used both as an objective of management and as a dependent/criterion variable to evaluate the effect of various factors on performance.

Despite its widespread use, ROI has been extensively criticized as being a totally inadequate indicator of economic rate of return (G. C. Harcourt, 1965; Ezra Solomon, 1971; Franklin Fisher and John McGowan, 1983). ROI, most typically defined as $Net\ Income_t/Total\ Assets_{t-1}$, does not properly relate the stream of profits to the investment that produced it. The earnings numerator is a consequence of investment decisions made in the past, but the assets denominator can be expected to have had influence on not only past and current earnings but also to have influence on future earnings. Because of this failure to produce an accurate mapping, ROI has been criticized as being so seriously flawed that it bears little, if any, resemblance to the crucial concept of internal or economic rate of return. The shortcomings of ROI are said to be so severe that its cross-sectional variations can be completely explained by the inappropriateness of the measure.

Due to a lack of proven validity, empirical investigations using ROI have been labeled by Fisher and McGowan as "totally misleading enterprises" and by George Benston (1985) as "of doubtful value." Those making use of accounting ROI argue that the noise created by the accounting distortions need not be expected to drown out the underlying signal of economic return contained in ROI and, therefore, ROI is still appropriate for use in analysis, (F. M. Scherer, 1979, and William Long and David Ravenscraft, 1984).

This paper examines the validity of ROI as a measure of economic rate of return by ascertaining the existence and extent of the association of corporate level accounting ROI with stock return, a widely accepted indicator of business performance. The null hypothesis for the analysis is that ROI has no validity as a measure of business performance and, as such, will not exhibit an association with stock return. While the tests provide direct evidence on the validity of corporate level ROI, the same underlying theorized inadequacies of accounting ROI are also present at the SBU level.[1] Conclu-

*School of Business Administration, University of Washington, Seattle, WA 98195. I am indebted to David Aaker, John Narver, and Eric Noreen for their help in various parts of this study. I especially thank Franklin Fisher for his invaluable comments and suggestions.

[1]A number of other issues relating to the validity of SBU profitability analysis, for example, the allocation of expenses and the setting of transfer prices, concern the validity of accounting data. These issues are not addressed in this study as they involve an entirely

sions regarding the validity of corporate level ROI may thus be largely extrapolated to the SBU level where pertinent stock price data are never available.

The findings of this analysis suggest that ROI is a useful, and perhaps best available, indicator of business performance. Significant associations between stock return and information contained in ROI are observed. Additional validity is evidenced by the fact that ROI contains information associated with strictly interfirm differences in stock return and unsystematic stock return. Consistent with the notion of efficient markets, a measure of unanticipated ROI (i.e., ROI not predictable based on its own past values) exhibits the strongest correlation with stock return. While the ROI measure unquestionably has serious limitations, the findings as a whole suggest that ROI provides information as to economic rates of return.

I. Testing Methodology

A "true" measure of economic return for a grouping of assets, such as a firm, is quite elusive. Stock price based returns may be the most useful indicator.[2] Market efficiency implies that the price of a stock reflects all available information relating to the profitability of the corporation. Favorable (unfavorable) news tends to result in an increase (decrease) in stock prices. One way to test whether the market perceives ROI to have a positive, negative, or inconsequential association with *economic* return is to assess its correlation with *stock* return. Fisher and McGowan assert that "accounting rates of return are useful only insofar as they yield information as to economic rates of return" (p. 82). Therefore, the usefulness/validity of accounting ROI can be tested through the determination of its covariance with stock market returns.[3]

This test is in the spirit of a stream of accounting research assessing the information content of accounting earnings. Ray Ball and Phillip Brown (1968), the seminal study in this area, found that firms with positive earnings changes (earnings being the numerator in the ROI measure) tend to have positive stock price changes and firms with negative earnings changes tend to have negative stock price changes. Other studies have since found that there is a correlation between the magnitude, as well as the sign, of earnings changes and stock price changes.[4] However, these studies are not directly applicable to the issue of the validity of ROI. The criticisms of ROI (in particular, the criticism that the measure fails to yield an accurate mapping between a given investment outlay and the earnings stream it generates) go above and beyond the limitations of accounting earnings not reflecting economic earnings (Fisher).

The stronger the correlation of ROI with stock return, the stronger the validity of ROI as a measure of business performance. As economic return differs from stock return, this is a stringent test of the validity of ROI. The efficient markets hypothesis implies that all relevant profitability information is already incorporated in the price of the stock. Therefore, a measure of economic return may not be in complete correspondence with stock return as some of this information is presumably already reflected in the price of stock. In addition, factors other than changes in the underlying profitability of the firm may influence stock return. For instance, stock prices exhibit substantial fluctuation associated with expectations of merger/buy-

different set of concerns than the validity of accounting ROI as a measure of business performance (Fisher, 1984).

[2]Gerald Salamon (1985) uses accounting data to develop an internal rate of return estimate. This estimate is conditional on assumptions made about growth rates, cash flow patterns, and the "representative" project.

[3]Perhaps it should be noted that this is not a test of whether the market responds to accounting announcements. The market has probably already incorporated any relevant information contained in annual ROI from other sources. This test involves ascertaining whether information contained in ROI is, or is correlated with, information that is used by investors to influence stock prices.

[4]An extensive discussion of this research is provided in Phillip Brown et al. (1985).

out activity. These movements in the price of the stock seem to have little to do with the economic return of the firm's operations. These differences in the nature of the return measures will lead to an understatement of the validity of ROI. Still, economic return and stock return appear to have enough in common to expect that any indicator of economic return should be associated with stock return.

II. Data

Two types of data are required to investigate the association of stock return with accounting ROI. Income statement and balance sheet information is required to obtain estimates of accounting ROI. So as to make comparisons, stock price data are needed for this same group of firms. Information of this type, for a large number of firms over time, is available through Standard and Poor's Compustat and the University of Chicago's Center for Research in Security Prices (CRSP) data files.

Compustat provides annual accounting information for twenty-year periods for industrial companies listed on the NYSE and AMEX and also for some nonindustrial companies. CRSP provides monthly stock price information for companies listed on the NYSE. The sample of companies used in the analysis consists of firms on which both Compustat and CRSP reported for the entire twenty-year period 1963–82. To ensure a correspondence between time periods, the additional requirement that firms have a December 31 fiscal year is also imposed. A total of 241 firms met these criteria and are utilized in the analysis.

Ball and Brown, who use a similar selection criterion, note that the selection process reduces the generality of the results. The sample is not representative of all firms as, for instance, it does not include young firms, firms that have failed, nor those not reported on by Compustat and CRSP. However, they indicate that the firms in the sample are of importance in their own right, and that some universality can be expected. The limitations of ROI as a measure of economic return are not unique to any particular group of firms. Although a host of attributes, for example, the firm's growth rate and configuration of net cash inflows and outflows, influences the extent of ROI's potential validity as a measure of economic return.

III. Empirical Investigation

A. *The Association of ROI with Stock Return*

Equation 1.1 of Table 1 reports the results of regressing annual stock return, calculated using monthly data, on accounting ROI.[5] As evidenced by the t-statistic of 9.62, the two measures of return are significantly associated. An interpretation of this association is that the market utilizes information concerning profit rate performance that is, to some extent, depicted by ROI. The small R^2 value of .020 indicates that the market is also utilizing a great deal of other information. The hypothesis that ROI is an unbiased estimate of stock return (i.e., the joint hypothesis of zero intercept and unit slope) can be rejected. Surprisingly, or perhaps coincidently, the hypothesis of a unit correspondence between stock return and ROI cannot be rejected. The coefficient value of .9471, indicating this correspondence, is only one-half of a standard deviation away from 1.00. Thus, while ROI is a biased estimate of stock return, the bias may be limited to a constant understatement.[6]

Fisher and Michael van Breda (1984) both comment that if the general shape of the benefit profile associated with an investment stays constant, then ROI and economic rate of return will vary together. This implies that a weak time-series association between the two measures for an individual firm may exist, but that across firms no such association would be present. While this is a possible interpretation of the above findings, this hypothesis is contradicted by additional

[5]Only 19 years of data are available from the 20 years of observations due to the use of Book Value of Assets in period $t-1$ in the denominator of ROI.

[6]Given the expected existence of positive present value of growth opportunities, ROI should be expected to understate stock return.

TABLE 1—ASSOCIATION OF STOCK RETURN WITH ACCOUNTING ROI

Equation	R^2	Number of Observations
1.1 $StkR_{it} = .072 + .947*AccR_{it} + \varepsilon_{it}$ (8.11) (9.62)	.020	4579
1.2 $stkr_{it} = -.0024 + 1.071*accr_{it} + \varepsilon_{it}$ (.59) (15.83)	.052	4579
1.3 $StkR_{it} = .0827 + 1.078*AccOIR_{it}$ (7.24) (8.28)		
$-1.265*AccNOER_{it} + \varepsilon_{it}$ (5.52)	.020	4579
1.4 $\log(1 + StkR_{it}) = .292 + .0523*\log(AccNI_{it})$ (8.84) (7.57)		
$-.0592*\log(AccA_{it-1}) + \varepsilon_{it}$ (8.03)	.014	4378[a]

Notes: $StkR$ = stock return; $stkr$ = unsystematic stock return; $AccR$ = accounting return on investment; $accr$ = unsystematic accounting return on investment; $AccOIR$ = accounting operating income return; $AccNOER$ = accounting nonoperating expense return; $AccNI$ = accounting net income; $AccA$ = accounting book value of assets. The t-statistics are shown in parentheses.

[a]The sample size is reduced as observations with negative net income could not be used in estimating the equation.

findings indicating significant associations solely on the basis of cross-sectional data. Separate regressions of stock return on ROI for each of the nineteen years produce positive coefficients, statistically significant at the 95 percent level, in fourteen of the years.[7] And of these, 13 are also significant at the 99 percent level. The average value for the coefficient reflecting the correspondence is 1.01 and the R^2 values average .05. While deficiencies in the measure might be able to explain all cross-sectional variation in ROI, this is not found to be so. ROI contains information about economic return that is associated with cross-sectional differences in stock return.

Much the same conclusions can be drawn from the analysis of unsystematic return, that is, the difference between total return and the expected return as indicated by economywide conditions and the *beta* of the firm. Equation 1.2 reports a regression of unsystematic stock return on unsystematic ROI.[8] A highly significant association is observed. As evidenced by the R^2 of .052, a higher correlation exists between the unsystematic returns than exists between the total returns. In addition, the hypothesis that unsystematic ROI is an unbiased indicator of unsystematic stock return cannot be rejected. This association suggests that ROI contains the especially important information about supra and infra normal (i.e., risk adjusted) profitability.

B. *Alternative Accounting Performance Measures*

While the above results strongly indicate that ROI does yield information as to eco-

[7]Of the remaining coefficients, two indicate a positive association at above the 75 percent level, two are essential zero, and the other is significantly negative.

[8]Unsystematic stock return and unsystematic accounting return measures are calculated as: $stkr_{it} = StkR_{it} - \beta_i^{sp}*StkR_{mt}$ and $accr_{it} = AccR_{it} - \beta_i^{a}*AccR_{mt}$, where $StkR_{it}$ and $AccR_{it}$ are the stock return and ROI of firm i in period t; $StkR_{mt}$ and $AccR_{mt}$ are the market stock return and average ROI for year t; and β_i^{sp} and β_i^{a} are estimates of the stock price *beta* and accounting *beta* for firm i. These estimates are obtained through a Stein estimation procedure, allowing for unequal variances, suggested by Bradley Efron and Carl Morris (1975).

nomic rates of return, it may not be the best indicator. A number of studies advocate and use accounting measures that attempt to limit the distortions caused by book depreciation expense. This estimate, which influences both the numerator and denominator in the ROI measure, is sometimes regarded as little more than an accounting artifact. The most commonly used alternative profitability measures are the growth rate in Operating Income and Operating Income/Sales, that is, a measure of profit margin or return on sales. Operating income differs from the net income measure used in calculating ROI in that it does not account for nonoperating expenses such as depreciation, interest expense, and taxes.

The growth rate in operating income is advocated as a means of reflecting earnings while adjusting for the size of the firm without making use of a potentially seriously flawed asset measure. But, of course, operating income may change due to a change in the level of assets and so not necessarily reflect movement in the underlying profit rate. Profit margin is perhaps the most widely advocated alternative profitability measure. It is felt by some to be a better indicator of monopoly profits as it is an approximation to the Lerner Index (Stephen Martin, 1984). The extent that profit margin approximates the Lerner Index is questionable as the measure fails to account for the cost of capital. Comparing profit margins across firms/industries, where conditions are drastically different, may say little about differences in rates of return.

Both operating income growth and profit margin are significantly correlated with stock return. However, their association with stock return, as indicated by R^2 values of .009 and .003, respectively, are smaller than that reported for ROI in equation 1.1 of Table 1. A nonnested hypothesis test suggested by Harold Hotelling (1940) indicates that the predictive power of operating income growth and profit margin are significantly less than that of ROI at the 95 and 99 percent confidence level, respectively. In fact, profit margin seems to be useful in explaining stock return only to the extent that it contains information that is better reflected in ROI. When stock return is regressed jointly on ROI and profit margin, the impact of profit margin is essential zero and highly insignificant.

Unlike the result reported in Long and Ravenscraft that all profit rate accounting measures are much the same, significant differences in the magnitude of association with stock return are observed for the most commonly used accounting measures of business performance. The empirical rationale for the superiority of ROI over the alternative measures can be explained by the fact that market participants appear to react in much the same fashion to the various components comprising ROI.

It might be hypothesized that ROI is not a homogeneous measure since its numerator is comprised of rather diverse income and expense components. These different components may reflect, or, for that matter, may not reflect, different types of information about business profitability. They may, therefore, be perceived differently by stock market participants. To test for heterogeneity, stock return is jointly regressed on the two seemingly most conceptually different components of ROI, that is, Operating Income Return and Nonoperating Expense Return. The results of this regression are reported in equation 1.3 of Table 1.

Both operating income return and nonoperating expense are significantly related to stock return. As evidenced by coefficients of 1.08 and -1.27, the information content of both items is perceived in much the same fashion by the market. The hypothesis that the two coefficients are the same in magnitude but of opposite sign cannot be rejected. This supports the practice of aggregating the components of Net Income to form the composite measure ROI.

The ROI measure can be hypothesized to be heterogeneous for another reason. Market participants may not react in the same manner to differences in ROI caused by differences in net income as opposed to differences in assets. A test of this hypothesis, assuming a multiplicative relationship, is provided by a regression of the logarithm of $(1+\text{stock return})$ on the logarithm of net income and the logarithm of the previous

period's book value of assets. Under the hypothesis that the association of stock return with ROI does in fact depend on the ratio of net income to assets, the coefficients will be of the same magnitude but of the opposite sign. Equation 1.4 reports this regression. Both coefficients are significant at above the 99.9 percent confidence level. The difference in the magnitude of the coefficients, if one exists, is small. The test that the sum of coefficients is zero yields a t-statistic of 1.86 (i.e., a sum of .0069 with a standard error of .0037). Market participants perceive the level of assets to be as important as the amount of net income. The results are very suggestive that it is indeed the ratio of net income to assets that is of relevance.

C. *The Role of Efficient Markets*

Despite the association observed between the return series, neither equations 1.1 or 1.2 in Table 1 can be expected to capture the full extent of the possible associations between the series. Neither equation properly depicts the notion of market efficiency. Presumably the market has already incorporated any predictable profit rate information contained in ROI into the price of the stock. Only unanticipated changes in profitability should exhibit a correlation with stock return. A measure of unanticipated ROI is the residual obtained after subtracting the return that can be predicted based upon past values of ROI. This transformation can be expected to remove the anticipated profit rate information contained in ROI. Equally important, this transformation may yield a stronger association between the series as it filters out noise caused by predictable, firm-specific factors (for example, certain accounting practices) that influence ROI but do not reflect underlying economic return. The effects of systematic differences in accounting practices should be dissipated to the extent that they can be anticipated.

The time-series behavior of ROI is well approximated by a first-order autoregressive process. A regression of ROI on ROI lagged one year produces an autoregressive coefficient estimate of .835, with a standard error of .009, and an R^2 of .66. A similar autoregressive process is found to characterize unsystematic return. A regression of unsystematic ROI on unsystematic ROI lagged one year produces an autoregressive coefficient of .898, with a standard error of .007, and an R^2 of .78. The residual errors from these first-order autoregressions are estimates of unanticipated ROI and unanticipated unsystematic ROI.

The regressions of stock return on unanticipated ROI, and unsystematic stock return on unanticipated unsystematic ROI, are reported in equations 2.1 and 2.2 of Table 2, respectively.[9] As evidenced by the R^2 of .074 and the t-statistic of 18.57, the association of unanticipated ROI with stock return is significantly stronger than with ROI.[10] The higher coefficient estimate (i.e., 3.19) indicates that the market tends to make a greater adjustment to profitability information reflected in unanticipated ROI than is reflected in ROI itself. Equation 1.1 masks the association of ROI with stock return by not separating anticipated from unanticipated ROI. Unlike unanticipated ROI, anticipated ROI has almost no association with stock return. A regression of stock return on anticipated ROI yields an R^2 of .0004. These same conclusions can be drawn from equation 2.2 about the association of unsystematic stock return with unsystematic ROI.

Strictly cross-sectional analysis also substantiates the role of unanticipated ROI. Separate regressions of stock return on unanticipated ROI produce positive coefficients, statistically significant at the 95 percent level, for each of the eighteen years. In seventeen of these eighteen years, the coefficients are also significant at the 99.99 per-

[9]Consistent with the widely held view of stock prices following a random walk, stock return is found to be serially uncorrelated. Therefore, stock return may just as appropriately be labeled unanticipated stock price or unanticipated stock return.

[10]While the growth rate in operating income already approximates white noise, unanticipated profit margin exhibits a stronger association with stock return than does profit margin. Still, the association of unanticipated profit margin with stock return, i.e., $R^2 = .028$, is significantly weaker than that for unanticipated ROI.

TABLE 2—ASSOCIATION OF STOCK RETURN WITH UNANTICIPATED ACCOUNTING ROI

Equation	R^2	Number of Observations
2.1 $StkR_{it} = .1367 + 3.19{*}UnAccR_{it} + \varepsilon_{it}$ (26.17) (18.57)	.074	4338
2.2 $stkr_{it} = .0036 + 3.829{*}unaccr_{it} + \varepsilon_{it}$ (.87) (26.76)	.149	4338
2.3 $StkR_{it} = .1263 + 3.514{*}UnAccR_{it}$ (23.91) (19.59)		
$+1.960{*}UnAccR_{it+1} + \varepsilon_{it}$ (11.34)	.106	4097
2.4 $stkr_{it} = -.0025 + 4.032{*}unaccr_{it}$ (.62) (28.23)		
$+1.414{*}unaccr_{it+1} + \varepsilon_{it}$ (10.08)	.169	4097

Notes: See Table 1. *UnAccR* = unanticipated accounting return on investment; and *unaccr* = unanticipated unsystematic accounting return on investment.

cent level. The R^2 values of these regressions average .16. Information contained in ROI is clearly associated with cross-sectional differences in profitability.

The notion of market efficiency suggests that stock market participants efficiently incorporate information about current and future business profitability. Assuming that market participants do not have a completely short-term preoccupation with current year's income, a strictly contemporaneous association should not be expected to exist between ROI and stock return. Stock return may be found to lead ROI, as the market has already made use of information that will later be reflected in ROI. A correlation between stock return and future ROI, given the slight likelihood of stock prices actually causing accounting ROI, would be additional evidence of the validity of ROI as an indicator of economic return.

Equation 2.3 of Table 2 reports a regression of stock return on contemporaneous and one year in the future residual ROI. Equation 2.4 does the same for unsystematic residual return. The same conclusions can be drawn from both equations. The significant associations observed for stock return leading ROI by one year, and the increases in explanatory power, suggest current stock prices do incorporate information about profitability that will later be reflected in next year's accounting ROI.

IV. Conclusions

Unquestionably, ROI has serious limitations as a measure of business performance. The relatively small correlation between stock return and information contained in ROI diminishes the validity of conclusions drawn from analyses making use of ROI. It is the degree of limitation that is at issue. Claims, such as those by Fisher and McGowan, that studies making use of accounting ROI are "totally misleading enterprises," due to the limitations of ROI as a measure of economic return, seem to be overstatements. The findings, especially taken as a whole, strongly suggest that ROI does contain information (albeit small) about economic rate of return.[11] It is observed that

[11] In addition to the ability of ROI to yield information about economic rates of return, another concern is that the measurement error in ROI may be correlated with explanatory factors of profitability. In particular, Salamon indicated that systematic differences in accounting practices that are based on firm size lead to different ROI levels. Using a conditional estimate of internal rate of return, Salamon found a significant, though weak, association between the estimated measurement error and total assets. My analysis fails to support the contention that systematic measurement error invalidates profitability studies. The association of total assets with the residual from equation 1.1, i.e., an estimate of the measurement error in ROI, and with the residual from equation 2.1, i.e., an estimate of the

(i) ROI is significantly correlated with stock return, both on a pooled time-series cross-sectional basis and on a strictly cross-sectional basis; (ii) ROI has a statistically greater association with stock return than commonly advocated alternative measures of profitability (i.e., operating income growth and profit margin); (iii) Consistent with the notion of efficient markets, contemporaneous and "one year in the future" unanticipated ROI has the strongest association with stock return.

The ROI measure contains, or is correlated with, information that stock market participants deem important as to profit performance. As such, it has a correspondence with a measure of economic rate of return. Whether it is due to the filtering of profit information that the market has already incorporated into the price of the stock or to the filtering of information that the market deems unimportant, an estimate of unanticipated ROI exhibits an even higher association with stock return. This consideration is critical in determining the full extent of the validity of ROI. The latter possibility indicates that the information content of ROI has been extracted and is fully depicted. The former scenario is consistent with ROI having more validity than indicated in the analysis. This scenario, as well as stock return not having an exact correspondence with economic return, suggests the reported correlations should be considered lower bound indications of the validity of ROI.

There is little basis for deciding how high a correlation should be before a measure can be considered an adequate proxy. However, the results begin to suggest that information contained in ROI can be extracted to produce a measure capable of providing insights into the profit performance of SBUs. Consistent with the view of Ira Horowitz (1984), the use of ROI, or a transformed profitability measure such as unanticipated unsystematic ROI, as one input into the evaluation of business unit profitability is clearly warranted.

measurement error in unanticipated ROI, is both small (R^2 values of .00037 and .00026, respectively) and insignificant.

REFERENCES

Ball, Ray and Brown, Phillip, "An Empirical Evaluation of Accounting Income Numbers," *Journal of Accounting Research*, Autumn 1968, *6*, 159–78.

Benston, George J., "The Validity of Profits-Structure Studies with Particular Reference to the FTC's Line of Business Data," *American Economic Review*, March 1985, *75*, 37–67.

Brown, Philip, Foster, George and Noreen, Eric, *Security Analyst Multi-Year Earnings Forecasts and the Capital Market*, Studies in Accounting Research No. 21, American Accounting Association, 1985.

Efron, Bradley and Morris, Carl, "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, June 1975, *70*, 269–77.

Fisher, Franklin M., "The Misuse of Accounting Rates of Return: Reply," *American Economic Review*, June 1984, *74*, 509–17.

______ **and McGowan, John J.,** "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, *73*, 82–97.

Harcourt, G. C., "The Accountant in a Golden Age," *Oxford Economic Papers*, March 1965, *17*, 66–80.

Horowitz, Ira, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, 74, 492–93.

Hotelling, Harold, "The Selection of Variables for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters," *Annals of Mathematical Statistics*, September 1940, *11*, 271–83.

Long, William F. and Ravenscraft, David J., "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 494–500.

Martin, Stephen, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 501–06.

Reese, James S. and Cool, William R., "Measuring Investment Center Performance," *Harvard Business Review*, May-June 1978, *56*, 28–46.

Salamon, Gerald L., "Accounting Rates of Return," *American Economic Review*, June 1985, *75*, 495–504.

Scherer, F. M., "Segmental Financial Reporting: Needs and Trade-Offs," in Harvey Goldschmid, ed., *Business Disclosure: Government's Need to Know*, New York: McGraw-Hill, 1979, 3–57.

Solomon, Ezra, "Return on Investment: The Relation of Book-Yield to True Yield," in J. Leslie Livingstone and Thomas J. Burns, eds., *Income Theory and Rate of Return*, Columbus: Ohio State University Press, 1971, 105–17.

Van Breda, Michael, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 507–08.

On the Comparative Statics of a Competitive Industry

By MICHAEL BRAULKE*

Except for rather unrealistic special cases, little is known about the comparative statics of a competitive industry in long-run equilibrium. This is surprising in view of the central role this concept plays in much of economic theory and policy. The likely reason is that the analysis of industry behavior in short-run equilibrium (where, by definition, the number and composition of firms in the industry remain fixed) is already quite complex.

Consider, for instance, some autonomous price change. Since this price change is likely to trigger other price adjustments at the industry level, the industry's overall response is composed of its reaction to the initial price change *and to the multiplicity of such echo effects*. The situation looks even more complicated in the long run when firms leave or enter and the composition of the industry changes. The final outcome of the adjustment process to a new short- or long-run equilibrium will depend fundamentally on how the industry's clientele and suppliers react. Hence, without relying on suitable assumptions concerning the reactions of the industry's market partners, virtually nothing can be said about its behavior in short- or long-run equilibrium.

Clear results can be obtained for an industry's *aggregate* response if one is willing to assume, as is commonly done, that the industry faces "normal conditions" in all its markets (where normal conditions mean essentially that demand schedules for the industry's outputs should not be rising or supply schedules for inputs should not be falling with their respective prices). Coldwell Daniel (1970), and Ronald Heiner (1982) with his more general approach, demonstrated that an industry's short-run equilibrium demand for inputs will conform to the traditional law of demand provided the industry faces normal conditions in its output market. Their analyses were restricted to the special case of a single-output industry in short-run equilibrium that faces a perfectly elastic supply in all its input markets. Yet their findings are also true for the general case of a multiproduct industry facing a less than perfectly elastic demand or supply in an arbitrary number of markets, as will be demonstrated below.

It is useful to distinguish between two concepts of long-run equilibrium when analyzing the industry's long-run equilibrium supply and demand behavior. The *narrow* concept defines long-run equilibrium as a state in which all firms in the industry realize zero profits so that all of them are in fact marginal. Consider the very special case of a single-output industry with identical firms and perfectly elastic supply in all input markets. The theory of the firm in long-run equilibrium developed by Charles Ferguson and Thomas Saving (1969), Lowell Bassett and Thomas Borcherding (1970), Daniel, Eugene Silberberg (1974b), and others suggests a full analogy to the findings on short-run industry behavior. Such an industry's long-run equilibrium demand for factors will also obey the law of demand provided normal conditions prevail in its output market.[1]

*Department of Economics, Universitaet Osnabrueck, D-4500 Osnabrueck, F.R.G. I gratefully acknowledge substantial help from Ron Heiner.

[1]As Silberberg has shown, the individual firm belonging to a competitive single-output industry must behave in long-run (and zero profit) equilibrium as if it were minimizing average costs, which in turn implies that its relative input demand (i.e., long-run input demand per unit of output) will fall if that input's price rises (1974b, Corollary 2). In the case just mentioned such an input price increase must eventually be passed on entirely to the output price. The industry's total output will consequently fall provided aggregate demand for it behaves normally. Any input price increase must therefore lead to a decline in the industry's long-run demand for that input since, with identical firms, aggregate input demand is nothing but the product of relative input demand of a representative firm and aggregate output.

This older literature is of little help, however, if a multiproduct industry is to be considered; if several markets are characterized by a less than perfectly elastic supply or demand; or if the exogenous price change originates in a less than perfectly elastic market. More important, the traditional theory of the firm in long-run equilibrium is not applicable to the *wide* concept of long-run equilibrium which more realistically allows for heterogeneous firms, hence the existence of inframarginal firms in long-run equilibrium. The existence of inframarginal firms appears to make the consequences of the entry and exit process much less intelligible. This has led previous authors to a rather pessimistic view on the predictability of aggregate industry behavior.

The following excerpt from Silberberg illuminates the apparent problem:

> There is no way to know a priori whether firms with the largest increases in AC will leave the industry first because these firms might also be the ones earning the largest rents. Because of this, it is not possible to identify how the total use of the factors in question by the industry will respond to a change in that factor's price. If a factor price p_i increases, the firms that are heavy users of x_i may expand relative to the other firms in the industry producing an upward sloping industry factor demand curve. ...Whether or not this situation is empirically relevant, it appears that downward sloping factor demands for an industry composed of nonidentical firms will have to remain an asserted rather than a derived result. [1974b, p. 740]

As will become evident below, such a pessimistic view is premature. Nor is it really necessary to follow in the footsteps of subsequent authors who directly or indirectly restricted the feasible range of diversity among firms in order to reach definite results on industry long-run behavior.[2] *The entry and exit process which seems to make everything so complicated follows a surprisingly simple inner logic. In fact, the "informational content" of this inner logic together with the usual assumption that normal conditions prevail in the industry's markets turn out to be sufficient to make its long-run behavior predictable.* It is the chief objective of this paper to demonstrate this and to do so without recourse to differentiability assumptions. Such differentiability assumptions have been very popular in the literature on industry behavior even though they are clearly inappropriate in view of the intrinsically discrete nature of the entry and exit process.

I begin by summarizing some useful aspects of the short-run supply and demand behavior of competitive multiproduct firms, and then define more precisely what I mean by normal conditions in the industry's markets. Subsequently, a basic result essentially due to Heiner is proved for an industry's short-run equilibrium behavior. Section II is devoted to the industry's long-run equilibrium behavior. It first discusses the implications of the entry and exit process and then presents a simple proof of the main result for long-run industry behavior that underlines its close relationship to Heiner's result for the short run.

I. The Industry's Short-Run Supply and Demand Behavior

Consider a competitive industry consisting of a set J of not necessarily identical profit-maximizing firms and let J^0 denote the subset of currently active firms. I define the short run to be too short for entries and exits to take place and hence identify an industry's short-run behavior simply with the sum of the responses of all active firms belonging to the frozen industry structure J^0.

Let q^j denote firm j's vector of output and input quantities and, following a useful convention, identify positive elements with outputs and negative elements with inputs. Furthermore, let the vector of markets prices p, which the firms face, be ordered conformably with the quantity vector q. Although some or all of these prices will have to be viewed as being determined at the industry

[2] As an example of the former, see John Panzar and Robert Willig (1978), and as an example for the latter, see Heiner (p. 560).

level by a market mechanism that equilibrates aggregate supply and demand, all of them are, of course, exogenous from the point of view of the individual firm. If $q^j(p)$ denotes firm j's profit-maximizing choice, its individual response matrix with respect to isolated price changes must consequently be positive semidefinite (and symmetric), or, in short,[3]

$$(1) \qquad q_p^j(p) \geqq 0.$$

This well-known property[4] comprises the traditional law of supply and demand for competitive firms,[5] but it says little about the behavior of an entire industry. Using capital letters to indicate summation over all active firms, we have

$$(2) \qquad Q(p) = \sum_{J^0} q^j(p),$$

so that (1) immediately implies the summed response to an isolated price change to be also positive semidefinite, that is,

$$(3) \qquad Q_p(p) = \sum_{J^0} q_p^j(p) \geqq 0.$$

It is, however, important to note that (3) is in general *not* identical with an industry's short-run equilibrium response to an exogenous price change unless this industry happens to face a perfectly elastic demand and supply in all its output and input markets. Yet, some of these markets may in fact be perfectly elastic while others may not. To find the industry's complete short-run equilibrium response, one has to account for the fact that any initially isolated price change may disturb the equilibrium in some of the industry's markets, and thus trigger further price adjustments.

In order to disentangle the exogenous from the endogenous component in the market prices p, it is useful to write

$$(4) \qquad p = P + \alpha,$$

where the vector α represents the truly exogenous components of prices such as taxes or subsidies, while the vector P stands for the components that are endogenous in the sense that prices have to clear the markets. While it is not the only conceivable interpretation, it is quite helpful to identify p, for the moment, with the prices which the firms in the industry face, and P with the prices (net of taxes, etc.) which the industry's clientele and suppliers actually pay or receive. Denote the aggregate demand of this clientele and the aggregate supply of its suppliers by the vector $X(P)$ where, again in line with the earlier adopted convention, output quantities are measured positively and input quantities negatively. Equilibrium in the industry's markets then requires $Q(p) - X(P) = 0$, or, using the decomposition specified in (4),

$$(5) \qquad Q(P+\alpha) - X(P) = 0.$$

Given the exogenous price components α, (5) can be viewed as determining implicitly the short-run equilibrium values $P^S = P^S(\alpha)$ of the endogenous price component. We are not interested in the question of how and whether the market reaches these equilibrium values, and hence simply assume their existence. Substituting this "solution" into (4) gives the short-run equilibrium prices $p^S(\alpha) = P^S(\alpha) + \alpha$, and substituting further into (2), yields the industry's short-run equilibrium supply and demand

$$(6) \qquad Q^S(\alpha) = Q\big(P^S(\alpha) + \alpha\big).$$

[3] To avoid notational clutter, partial derivatives with respect to an argument are denoted by subscripts. Thus, with q and p being (column) vectors of equal order, q_p is the quadratic matrix with the typical element $\partial q_i / \partial p_j$. Also, I will identify the symbols $\geqq 0$ and $\leqq 0$ with positive and negative semidefiniteness, respectively, if quadratic matrices are involved. Furthermore, a vector product such as pq should always be read to mean an inner product. Thus, pq denotes a firm's profit.

[4] See, for example, Silberberg (1974a). Positive semidefiniteness and symmetry of the response matrix q_p is most easily derived as a property of the indirect profit function.

[5] Positive semidefiniteness of q_p implies in particular $\partial q_i / \partial p_i \geqq 0$ for all i. Remembering that inputs are measured negatively, this means that the firm's supply of an output (or demand for an input) does not fall (increase) with a rise in its own-price.

Q^S depends only on the exogenous price component α, since the endogenous component P always has to adjust, so as to clear the markets.

So far it has not been specified how aggregate demand and supply in the industry's markets behave. The industry will be said to face "normal conditions" in its markets if the Jacobian of the demand and supply functions is negative semidefinite, that is, if

$$(7) \qquad X_P(P) \leqq 0$$

holds. Hardly any justification is required to identify normal behavior of aggregate demand for the industry's outputs and supply of its inputs with this property (7) as long as these demands and supplies are not interrelated; in this case the Jacobian is diagonal with all the (nonpositive) own-price effects on the main diagonal and hence clearly negative semidefinite. If these markets are, however, interconnected through the prices P, then (7) is a natural characterization of what is intuitively meant by normal conditions. Indeed, (7) guarantees that no autonomous increase in the industry's supply of an output or decline in its demand for an input can ever lead to a rise in the corresponding short-run equilibrium price of that particular output or input.[6] Technically speaking, (7) ensures that own-price effects outweigh all cross repercussions. I should add here that for my present purposes, I could equally well use an almost identical definition of normal conditions which merely requires that, given any change in prices from P^0 to P^1, the corresponding changes in demand and supply be nonpositively correlated, that is, that

$$(8) \quad (P^1 - P^0)\left(X(P^1) - X(P^0)\right) \leqq 0$$

holds. That (7) implies (8) is easily seen when applying Taylor's theorem on the vector product $(P^1 - P^0)(X(P^1) - X(P^0))$ and observing that it may be expressed[7] as a quadratic form in X_P which must be nonpositive by (7). A converse line of reasoning may not be feasible particularly if $X(P)$ is not differentiable everywhere. However, I maintain both these definitions of normality side by side because (7) is more clearly related to the partial-equilibrium tradition and (8) is all that is needed to prove the following

PROPOSITION 1 (Heiner): *Let the exogenous price components change from α^0 to α^1 and the industry's aggregate short-run equilibrium supply and demand change correspondingly from $Q^S(\alpha^0)$ to $Q^S(\alpha^1)$. If the industry faces normal conditions in all its markets, these changes will be nonnegatively correlated, that is,*

$$(\alpha^1 - \alpha^0)\left(Q^S(\alpha^1) - Q^S(\alpha^0)\right) \geqq 0.$$

PROOF:

Let $p^0 = P^0 + \alpha^0$ and $p^1 = P^1 + \alpha^1$ denote the short-run equilibrium prices when α^0 or α^1 prevail, respectively, and write q^0 and q^1 for the corresponding profit-maximizing input-output choices of the firms.[8] Since maximal profits cannot be less than profits with a feasible but not necessarily optimal choice of inputs and outputs, we have

$$\begin{aligned}\sum_{J^0} p^0 q^0 - \sum_{J^0} p^0 q^1 &= p^0\Big(\sum_{J^0} q^0 - \sum_{J^0} q^1\Big)\\ &= p^0\left(Q^S(\alpha^0) - Q^S(\alpha^1)\right)\\ &\geqq 0\end{aligned}$$

[6] Consider an equilibrium situation described by $Q(P^* + \alpha) - X(P^*) = 0$ and suppose there is an autonomous increase (decline) in the industry's supply of output i (demand for input i) so that $dQ_i > 0$ and $dQ_{j\neq i} = 0$. The consequent change in equilibrium prices must then solve $Q_p dP^* + dQ - X_P dP^* = 0$ or $dQ = (X_P - Q_p)dP^*$. Multiply from left with dP^* and note that by construction dP^*dQ reduces to $dP_i^* dQ_i$. Hence, $dP_i^* dQ_i = dP^*(X_P - Q_p)dP^*$. Consequently, $X_P - Q_p \leqq 0$ is a necessary, and $X_P \leqq 0$ is in view of (3), a sufficient condition to guarantee that dp_i^* $(= dP_i^*)$ is indeed nonpositive.

[7] Define the *scalar* function $f(P) = (P^1 - P^0)X(P)$ and note that by Taylor's theorem we may write $f(P^1) = (P^1 - P^0)X(P^1) = (P^1 - P^0)X(P^0) + (P^1 - P^0)X_P(\cdot)(P^1 - P^0)$, where X_P has to be evaluated at a suitable point in the range between P^0 and P^1. Subtracting $(P^1 - P^0)X(P^0)$ we have $(P^1 - P^0)(X(P^1) - X(P^0)) = (P^1 - P^0)X_P(\cdot)(P^1 - P^0)$, a quadratic form in X_P.

[8] From now on I drop the firm-specific superscript j without dropping, however, the assumption that firms may be as unequal as they wish.

and, conversely,

$$\sum_{J^0} p^1 q^1 - \sum_{J^0} p^1 q^0 = p^1\big(Q^S(\alpha^1) - Q^S(\alpha^0)\big)$$
$$\geqq 0.$$

Adding these two inequalities, substituting $\alpha^1 - \alpha^0 + P^1 - P^0$ for $p^1 - p^0$, and rearranging slightly gives

$$(\alpha^1 - \alpha^0)\big(Q^S(\alpha^1) - Q^S(\alpha^0)\big)$$
$$\geqq -(P^1 - P^0)\big(Q^S(\alpha^1) - Q^S(\alpha^0)\big).$$

Since all markets have to clear, the right-hand side equals $-(P^1 - P^0)(X(P^1) - X(P^0))$ and is hence nonnegative in view of (8) or, as demonstrated above, also by (7).

To avoid repetition, I postpone the interpretation of this result until after Proposition 2 which characterizes an industry's long-run behavior.

II. The Industry's Long-Run Supply and Demand Behavior

The analysis of the industry's long-run supply and demand behavior can be carried out much along the lines of the previous section, yet it differs in one essential aspect: the set of actively participating firms in the industry (so far denoted by J^0) is no longer fixed but determined in long-run equilibrium by profitability and is hence the result of an entry and exit process. We must therefore, if only briefly, first investigate the essential characteristics of this process which may take place between two long-run equilibria.

Consider again the transition in the exogenous price components from α^0 to α^1, let $p^0 = P^0 + \alpha^0$ and $p^1 = P^1 + \alpha^1$ denote the corresponding long-run equilibrium prices and write J^0 for the initial and J^1 for the final long-run equilibrium configuration of active firms. Furthermore, let q^0 and q^1 denote the profit-maximizing output and input choices when prices p^0 and p^1, respectively, prevail, and the always feasible alternative to withdraw and produce nothing $(q = 0)$ were ruled out. Since zero profits form the dividing line in the decision to enter or exit, it is easy to see that in the transition from J^0 to the new industry structure J^1,

$$(p^1 - p^0)q^1 \geqq 0 \tag{9a}$$

must hold for all entering firms, whereas

$$(p^1 - p^0)q^0 \leqq 0 \tag{9b}$$

must hold for all exiting firms.[9] Thus, the quantities which entering firms add to aggregate industry supply and demand are positively correlated, and the quantities which the exiting firms, if any, withdraw from aggregate supply and demand are negatively correlated with the changes in long-run equilibrium prices. Using this characterization of the entry and exit process as well as the usual information on how the firms remaining in the industry respond to price changes, it would be possible (but tedious) to determine the industry's long-run equilibrium response. A much more direct approach is to start with the following two sets of industry inequalities. First note that

$$\sum_{J^1} p^1 q^1 \geqq \sum_{J^0} p^1 q^1 \geqq \sum_{J^0} p^1 q^0 \tag{10}$$

must hold. The first inequality sign in (10) reflects the entry and exit process. By its very nature it guarantees that the long-run equilibrium configuration J^1 contains all and at the same time exclusively firms that are able to break even at the going equilibrium prices p^1 whereas the configuration J^0 may contain firms that incur losses at these prices. Hence, total profits of the equilibrium configuration J^1 at prices p^1 cannot be smaller than total profits associated with any other configuration. The second inequality sign rests on the conventional argument that maximal profits are never smaller than profits with a possibly inappropriate input/output choice. For entirely analogous reasons,

[9] Observe that for entering firms $p^1q^1 \geqq 0 \geqq p^0q^0$ must hold because, otherwise, they would have no reason to enter. Since furthermore $p^0q^0 \geqq p^0q^1$ by the conventional profit maximum property, (9a) follows. The argument for (9b) runs symmetrically.

the following set of inequalities must hold,

$$(10') \quad \sum_{J^0} p^0 q^0 \geqq \sum_{J^1} p^0 q^0 \geqq \sum_{J^1} p^0 q^1.$$

Now, adding (10) and (10′), writing $Q^L(\alpha^0) = \Sigma_{J^0} q^0$ and $Q^L(\alpha^1) = \Sigma_{J^1} q^1$ and rearranging slightly, we have

$$(11) \quad (p^1 - p^0)(Q^L(\alpha^1) - Q^L(\alpha^0)) \geqq 0,$$

which is the basis for

PROPOSITION 2: *Let the exogenous price components change from α^0 to α^1 and the industry's aggregate long-run equilibrium supply and demand change correspondingly from $Q^L(\alpha^0)$ to $Q^L(\alpha^1)$. If the industry faces normal conditions in all its markets, these changes will be nonnegatively correlated, that is,*

$$(\alpha^1 - \alpha^0)(Q^L(\alpha^1) - Q^L(\alpha^0)) \geqq 0.$$

PROOF:

Starting from (11), proceed as in the proof of Proposition 1.

Propositions 1 and 2 have two major implications:

(i) As long as normal conditions prevail in an industry's markets, its aggregate supply and demand will obey the traditional law of supply and demand in both the short and long run. This is easily seen when fixing all but the ith exogenous price component and observing that the inequalities stated in the two propositions collapse to $(\alpha_i^1 - \alpha_i^0)(Q_i^t(\alpha^1) - Q_i^t(\alpha^0)) \geqq 0$ with $t = S, L$, and i specifying any output or input. As a change in α_i may represent either a change in an indirect tax or subsidy or simply a (vertical) shift in the supply or demand schedules in any of the industry's markets, this result is quite comprehensive. In particular, it states that an indirect tax or a subsidy levied on some input or output will never be counterproductive at the industry level.[10] Similarly, an autonomous rise in the demand for one of the industry's outputs or in the supply of one of its inputs can never lead to a fall in the industry's supply of that output or demand for that input. Note that all these conclusions hold irrespective of whether the particular market under consideration is characterized by perfectly or less than perfectly elastic demand or supply, and regardless of how dissimilar (heterogeneous) individual firms might be from each other.

As can be seen from the proof, the normality assumption is sufficient but not necessary for the validity of the law of supply and demand. As a consequence, small deviations from negative semidefiniteness of the Jacobian X_p need not topple the law of supply and demand, but sufficiently large ones will definitely do so even if all own-price effects in X_p have the correct sign.

(ii) What must hold for the entire industry, given normal conditions in its markets, need *not* necessarily hold for the individual firm. Indeed, unless we can assume for the short-run analysis that firms are identical, the overall short-run response of some firms may differ qualitatively in direction from that of the entire industry.[11] The same applies for the individual firms' long-run equilibrium response even if I were to assume essentially identical firms (which I don't). The reason is that I am unable to determine how the number of firms in the industry develops. When a particular price component α_i changes, the number of firms may increase or decrease from one initial long-run equilibrium to the next, and it is therefore quite possible that an individual firm's long-run response is *opposite* to that of the entire industry.[12]

It may be tempting to ask whether something definite can be said about the relative

[10] Compare William Baumol and Wallace Oates, who in the context of an environmental tax, raised doubts as to its long-run effectiveness (1975, ch. 12, Proposition 4). However, in a full general equilibrium context, there would still remain an ambiguity associated with how the tax revenue is spent.

[11] This may not be obvious from the analysis presented here. See Heiner or my earlier paper (1984) for more details.

[12] For an interesting discussion of the consequences of this possibility for the effects of an environmental tax on pollution, see Alfred Endres (1983).

strength of an industry's long-run response compared to its short-run equilibrium response by analogy to phenomena resting on the strong LeChâtelier principle (which are well-known from the traditional theory of the competitive firm). However, no such systematic relationship appears to exist, at least as long as we consider finite changes in the parameters α. This should hardly be surprising since the strong LeChâtelier principle is also a strictly local phenomenon that cannot be expected to hold globally. Even if I were to consider only marginal changes in the parameters α, it would be out of the question to force the indispensable differentiability assumption on $Q^L(\alpha)$ since, due to the intrinsically discrete entry and exit process, at least the long-run equilibrium supply and demand may in fact have jump discontinuities at any point α.

III. Summary

I have shown that the basic law of supply and demand holds for the aggregate output-supply and input-demand behavior of a competitive industry in *both* the short and long run, provided the industry faces normal conditions in its product and factor markets (i.e., negatively sloped output-demand and positively sloped input-supply and no or sufficiently weak cross-price relations). These results hold irrespective of whether the individual firms are in any way similar or different from each other (except, of course, for their profit-maximizing behavior). On the other hand, these results *cannot* be obtained by aggregating over individual firms acting in isolation from each other: any exogenous price change is likely to trigger further price adjustments and in the long run also entries and exits, and these repercussions inextricably cloud the individual firm's overall response.

Yet it turns out ultimately to be immaterial as to what the consequent adjustments in the equilibrium prices might be or how the configuration of active firms in the industry might develop. The apparent ambiguity can be eliminated by recourse to the assumed normal behavior of the industry's market partners and by recognizing that the entry and exit process guarantees that the net addition to industry supply and demand must always be nonnegatively correlated with the changes in long-run equilibrium prices.

REFERENCES

Bassett, Lowell R. and Borcherding, Thomas E., "Industry Factor Demand," *Western Economic Journal*, September 1970, *8*, 259–61.

Baumol, William J. and Oates, Wallace E., *The Theory of Environmental Policy*, Englewood Cliffs: Prentice Hall, 1975.

Braulke, Michael, "The Firm in Short-Run Industry Equilibrium: Comment," *American Economic Review*, September 1984, *74*, 750–53.

Daniel, Coldwell III, *Mathematical Models in Microeconomics*, Boston: Allyn and Bacon, 1970.

Endres, Alfred, "Do Effluent Charges (Always) Reduce Environmental Damage?," *Oxford Economic Papers*, July 1983, *35*, 254–61.

Ferguson, Charles E. and Saving, Thomas R., "Long-Run Scale Adjustments of a Perfectly Competitive Firm and Industry," *American Economic Review*, December 1969, *59*, 774–83.

Heiner, Ronald A., "Theory of the Firm in 'Short-Run' Industry Equilibrium," *American Economic Review*, June 1982, *72*, 555–62.

Panzar, John C. and Willig, Robert D., "On the Comparative Statics of a Competitive Industry with Inframarginal Firms," *American Economic Review*, June 1978, *68*, 474–78.

Silberberg, Eugene, (1974a), "A Revision of Comparative Statics in Economics, or, How to Do Comparative Statics on the Back of an Envelope," *Journal of Economic Theory*, February 1974, *7*, 159–72.

______, (1974b), "The Theory of the Firm in 'Long-Run' Equilibrium," *American Economic Review*, September 1974, *64*, 734–41.

The Elasticity of Scale, the Shape of Average Costs, and the Envelope Theorem

By CHARLES F. REVIER*

In a widely cited note in this *Review* (1975), Giora Hanoch drew attention to the distinction between two different concepts of returns to scale, one concerning the relative change in output for equiproportionate changes in all inputs along a ray from the origin, and the other concerning the change in output relative to costs along the expansion path. Hanoch demonstrated that the two concepts give equal measures for the point-elasticity of scale, ε, at any point on the expansion path for any production function. However, if the production function is nonhomothetic, the *rate of change* in ε with output along a ray is generally *not* equal to its rate of change along the expansion path. Hanoch also demonstrated that the shape of the average cost curve depends upon the change in ε along the expansion path, not along a ray, and that the assumption of a downward-sloping technically optimal surface (where $\varepsilon = 1$), with $\varepsilon < 1$ above it and $\varepsilon > 1$ below it, is neither necessary nor sufficient for classical U-shaped average cost curves.

This note utilizes the envelope theorem to provide an alternative derivation of Hanoch's results concerning the relationship between changes in ε along a ray and changes in ε along the expansion path. This approach gives greater intuitive insight into Hanoch's conclusions. It allows graphical illustration of the results, with obvious pedagogical rewards. In addition, the derivation provides one significant new result, namely, a general proposition that the rate of change in ε along a ray must be algebraically less than its rate of change along the expansion path at a point where the ray and the expansion path intersect.

I. The Equality of the Two Measures of Returns to Scale

The elasticity of output with respect to (minimized) cost at a point along the expansion path is the ratio LAC/LMC, where LAC is long-run average cost and LMC is long-run marginal cost. The elasticity of output with respect to an equiproportionate change in all inputs may be defined as the ratio of the percentage change in output to the percentage change in cost along a ray from the origin, since the percentage change in this "ray cost" must be the same as the equal percentage change in all the inputs (assuming fixed input prices).[1]

The long-run total cost function, relating minimized cost to output along the expansion path, is the indirect objective function for the following optimization problem:

$$(1) \qquad \min_{\{x_1,\ldots,x_n\}} \sum_1^n p_i x_i$$

$$\text{subject to} \qquad f(x) = y,$$

where $x = (x_1, \ldots, x_n)$ is the vector of input quantities, $p = (p_1, \ldots, p_n)$ is the vector of input prices, $f(x)$ is the production function, and y is some given output level. Now consider adding the constraint that the inputs be used in fixed proportions, namely, the proportions of the input quantities that solve the original problem (1). In other words, this additional constraint is just binding at the original optimum. The new optimization problem with the added constraint

*Department of Economics, Colorado State University, Fort Collins, CO 80523. I am grateful to Robert R. Keller for helpful comments. I have also benefited from the comments of Liang-Shing Fan, Rodney D. Peterson, and George F. Rhodes on earlier drafts.

[1] This is established in Hanoch's equation (4), p. 493.

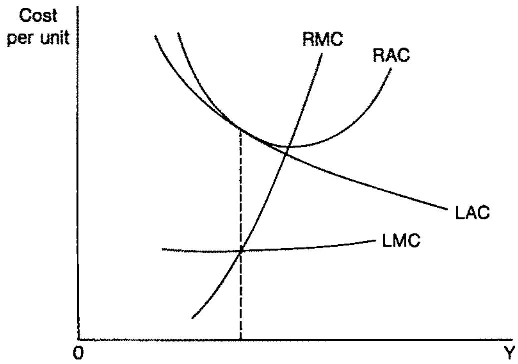


FIGURE 1

is actually a degenerate case, since normally there will be only one unique x which satisfies the constraints. However, it remains true that the indirect objective function for this degenerate optimization problem is a cost function—namely, the function giving the ray cost, or cost of producing a given output along the ray which maintains the initial input proportions. Dividing this ray cost by the quantity of output gives the ray average cost, RAC, and the rate of change of total ray cost with respect to output is the ray marginal cost, RMC. The elasticity of output with respect to an equiproportionate change in all inputs along the ray is therefore RAC/RMC.

At the original optimum point where the expansion path and the ray from the origin intersect, the ray cost and the (minimized) long-run total cost are identical, and therefore $LAC = RAC$. Furthermore, at any output level other than that at the original optimum point, the more-constrained total ray cost can never be less than the less-constrained long-run total cost. Therefore, these two total cost functions are tangent at the original optimum point, and their derivatives with respect to the output parameter must be equal at that point, that is, $RMC = LMC$.[2] Therefore

$$(2)\qquad LAC/LMC = \varepsilon = RAC/RMC.$$

At a point on the expansion path, the two concepts of returns to scale give identical values for the point-elasticity of scale.

Graphically, the long-run total cost curve is the envelope for the family of more-constrained total ray cost curves. Similarly, the LAC curve is the envelope for the family of RAC curves, as shown in Figure 1.

II. The Unequal Rates of Change of ε Along the Ray and Along the Expansion Path

The rate of change of ε with y along the expansion path is

$$(3)\qquad \left.\frac{\partial \varepsilon}{\partial y}\right|_p = \frac{\partial}{\partial y}\left(\frac{LAC}{LMC}\right)$$

$$= \frac{1}{LMC}\left(\frac{\partial LAC}{\partial y} - \varepsilon\frac{\partial LMC}{\partial y}\right).$$

Similarly, the change in ε along the ray is

$$(4)\qquad \left.\frac{\partial \varepsilon}{\partial y}\right|_x = \frac{\partial}{\partial y}\left(\frac{RAC}{RMC}\right)$$

$$= \frac{1}{RMC}\left(\frac{\partial RAC}{\partial y} - \varepsilon\frac{\partial RMC}{\partial y}\right).$$

At the point of intersection of the ray and the expansion path, $RMC = LMC$, and $\partial LAC/\partial y = \partial RAC/\partial y$ because of the tangency of the RAC curve and the LAC curve at that output level. However, $\partial LMC/\partial y \neq \partial RMC/\partial y$. In fact, since the total ray cost and long-run total cost functions are tangent with total ray cost never less than long-run total cost, it must be true that $\partial RMC/\partial y \geq \partial LMC/\partial y$ (the Le Chatelier Principle), and equality would hold only if the ray and the expansion path coincide over a range of output levels. For the regular case in which the ray intersects the expansion path at a single point, it must therefore be

[2] For a full discussion of these generalized envelope theorem results, see Eugene Silberberg (1971) and the extension in Silberberg (1978, pp. 293–98), to the case in which the relevant parameter appears in the constraints as well as in the objective function.

true that

$$(5) \qquad \partial\varepsilon/\partial y|_x < \partial\varepsilon/\partial y|_p.$$

Thus, ε increases at a faster rate or decreases at a slower rate along the expansion path than it does along the ray. This relationship is not apparent from Hanoch's derivation.

In his Proposition 1, part (ii), Hanoch establishes that if the production function is strongly quasi concave, then at a point of locally constant returns to scale where $\varepsilon=1$ it follows that if $(\partial\varepsilon/\partial y)|_p<0$ then $(\partial\varepsilon/\partial y)|_x<0$. This follows immediately from (5) above.

III. The Irrelevance of the Change in ε Along the Ray for the Shape of the LAC Curve

Hanoch's Proposition 1, part (iii), and Corollary 1, part (iii), state that if ε is decreasing through 1 along a ray, this is not a sufficient condition for the *LAC* curve to be U-shaped. In this case, *RAC* reaches a minimum at $\varepsilon=1$, but *LAC* may have a maximum, not a minimum, at this point. Hanoch proves this result by example. From the envelope theorem perspective, this possibility becomes intuitively clear. An *LAC* curve with an inverted U shape can very well be the envelope curve for a family of U-shaped *RAC* curves, as shown in Figure 2. The value of ε increases from values less than 1 to values greater than 1 along the expansion path, but its value decreases through 1 along any ray.

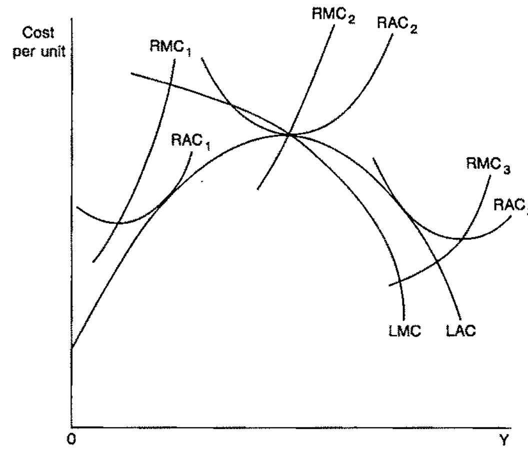


FIGURE 2

IV. Conclusion

Thus, viewing the *LAC* curve as the envelope for the family of curves showing average cost of production along each ray from the origin makes it much easier to perceive the relationship between returns to scale and the shape of the *LAC* curve. Moreover, the envelope approach directly establishes the new result that at any point along the expansion path the elasticity of scale never increases faster or decreases more slowly along the ray than it does along the expansion path.

REFERENCES

Hanoch, Giora, "The Elasticity of Scale and the Shape of Average Costs," *American Economic Review*, June 1975, *65*, 492–97.

Silberberg, Eugene, "The Le Chatelier Principle as a Corollary to a Generalized Envelope Theorem," *Journal of Economic Theory*, June 1971, *3*, 146–55.

______, *The Structure of Economics: A Mathematical Analysis*, New York: McGraw-Hill, 1978.

Unemployment as a Discipline Device with Heterogeneous Labor

By JON STRAND*

In a recent article (1984), Carl Shapiro and Joseph Stiglitz (S-S) present a model of equilibrium unemployment, where the threat of being fired acts as a deterrent against shirking among workers, assuming homogeneous labor, no firm reputation building, and no bonding of workers. In this paper, I extend the S-S model to the heterogeneous labor case, by assuming that some of the workers always have an incentive to shirk and will be fired when caught shirking. Among unemployed workers, some quit their last job voluntarily, and some were fired. I show that when nonshirkers are alike in my extended model and the S-S model, equilibrium will imply lower unemployment among nonshirkers when screening of applicants is efficient, and imply higher unemployment when screening is inefficient, in a sense to be defined below. When screening is inefficient, the dominating added effect is the lower average productivity among new hirees, resulting from some of those employed actually being shirkers, which makes employers less willing to hire at a given wage. With efficient screening, the most important new effect is the lowering of the wage necessary to deter good workers from shirking, since these now would be identified as bad if found shirking on the job, and would thereafter suffer a long period of unemployment. This lower wage makes the equilibrium labor force of each firm larger than in the S-S model. If the share of shirkers is sufficiently close to zero, this latter effect will always dominate and unemployment is always lower in my model.

Since the case exposed in the S-S model always entails underemployment, the type of labor heterogeneity introduced here may thus tend to make equilibrium either more or less efficient, depending on the efficiency of screening, and always more efficient when there are few bad workers. With perfect screening it may now even be possible for a competitive market to support a first-best equilibrium, with no unemployment at all among good workers, something that was never possible in the S-S model.

I. The Model

Assume that workers' productivities can take the values 1 (working) and 0 (shirking). Type 1 workers are shirkers by always choosing 0. The effort of type 2 workers is $e>0$ when choosing 1, and 0 when choosing 0. The share of type 1 workers is k, all live forever and discount at the rate r, and quit their jobs when employed at a rate b. The firm monitors its workers at a rate q. Shirkers are upon a control identified as being of type 1 and fired. The job acquisition rates of unemployed persons who quit and were fired from their last job, respectively, are a and γa, $\gamma \in [0,1]$, where lower levels of γ (assumed exogenous) imply more efficient screening. Define $V_{iE}(S)$ and $V_{iE}(N)$ as the expected discounted lifetime utilities of an employed type i worker who shirks and does not shirk, and $V_{iu}(Q)$ and $V_{iu}(F)$ as the corresponding utilities for unemployed workers who last quit and were fired, respectively, for $i=1,2$. We then have

$$(1)\quad rV_{2E}(S)=w+b\left[V_{2u}(Q)-V_{2E}(S)\right]+q\left[V_{2u}(F)-V_{2E}(S)\right],$$

$$(2)\quad rV_{2E}(N)=w-e+b\left[V_{2u}(Q)-V_{2E}(N)\right],$$

$$(3)\quad rV_{2u}(Q)=s+a\left[V_{2E}(N)-V_{2u}(Q)\right],$$

$$(4)\quad rV_{2u}(F)=s+\gamma a\left[V_{2E}(N)-V_{2u}(F)\right],$$

*Department of Economics, University of Oslo, Box 1095, Blindern, 0317 Oslo 3, Norway. I thank Gunnar Bramness, Kalle Moene, Joseph Stiglitz, seminar participants at Oslo and Stanford, and a referee for helpful comments.

where w is the wage paid to all and s is unemployed workers' utility. In each of these equations the left-hand side expresses the interest rate times the lifetime asset value of labor market participation which must equal the flow benefits plus expected change in asset value on the right-hand side.[1]

The system (1)–(4) can be solved for the endogenous variables $V_{2E}(S)$, $V_{2E}(N)$, $V_{2u}(Q)$, and $V_{2u}(F)$, for given levels of w and a. I am here only interested in the nonshirking condition (NSC) that must be fulfilled for type 2 workers not to shirk, that is, for $V_{2E}(N) \geq V_{2E}(S)$. This turns out to be

$$(5) \qquad w \geq s + \left(1 + \frac{r+b+a}{q} \cdot \frac{r+\gamma a}{r+a}\right)e.$$

With $\gamma = 1$, (5) is identical to the NSC in Shapiro-Stiglitz. With $\gamma < 1$, $\hat{w}$, defined by equality in (5), is now lower, as a result of screening by hiring firms. With $\gamma = 0$, screening is precise in the sense that firms know with certainty whether a worker about to be hired either quit or was fired from his last job. Only workers who quit will then be hired, and fired workers remain unemployed.

Firms are assumed to have decreasing returns production functions of the type (normalizing the number of firms to one)

$$(6) \qquad x = f(L_2), \quad f'(\cdot) > 0, \quad f''(\cdot) < 0.$$

The firms act competitively and maximize profits with respect to the number of workers they hire. The rate of monitoring q, and the precision of screening γ, are assumed exogenous to firms, that is, not subject to explicit maximization in the model.

[1]For example, in (1), $V_{2u}(Q) - V_{2E}(S)$ is the change in asset value from quitting, which occurs at a rate b, while $V_{2u}(F) - V_{2E}(S)$ is the corresponding change in value when the worker is fired, which occurs at the rate of monitoring, q, when the worker chooses to shirk. Mathematically, (1)–(4) can be derived considering the expected realized asset value over a small discrete period $[o, t]$, and taking the limit as $t \to 0$, as is done in Shapiro-Stiglitz. Note that it is assumed in (3)–(4) that unemployed type 2 workers are rationally expecting not to shirk, once they are again employed.

II. Market Equilibrium

To derive market equilibrium, define N as the fixed total labor supply, L as employment, and $U = N - L$ as unemployment, with subscripts i for the two types, such that $N_1 = kN$, $N_2 = (1-k)N$. Let $U_1(Q)$ and $U_1(F)$ be the number of unemployed type 1 workers who last quit and were fired, such that $U_1(Q) + U_1(F) = U_1$. For type 2 workers, the number of quitters must at equilibrium equal the number hired, that is,

$$(7) \qquad bL_2 = a(N_2 - L_2) \equiv aU_2.$$

For quitting and fired type 1 workers, we have in the same way $bL_1 = aU_1(Q)$ and $qL_1 = \gamma a \cdot U_1(F)$, and thus

$$(8) \qquad U_1 = (b/a + q/\gamma a)L_1.$$

Using that $L_i = N_i - U_i$ for $i = 1,2$, aggregate employment is now easily derived from (7)–(8), as

$$(9) \quad L = L_1 + L_2 = \left[\frac{\gamma ka}{q + \gamma(a+b)} + \frac{(1-k)a}{a+b}\right]N,$$

while the unemployment rate is

$$(10) \quad u = \frac{N-L}{N} = \frac{ka+b}{a+b} - \frac{\gamma ka}{\gamma(a+b)+q}.$$

The rate of hiring in the economy, H, must equal the rate at which workers leave jobs in equilibrium, that is, we must have

$$(11) \quad H = a[U_1(Q) + \gamma U_1(F) + U_2] = bL + qL_1.$$

The average productivity of hired workers now equals the share of type 2 workers among those hired, h, given by[2]

$$(12) \quad h = bL_2/(bL + qL_1) = \frac{(1-k)\gamma b(a+b) + (1-k)qb}{(b+qk)\gamma(a+b) + (1-k)qb}.$$

[2]The h is what Leif Johansen (1982) calls "efficiency in."

The expected present discounted value of output Q, produced by an additional hired worker, equals h times the value of a type 2 worker, that is,

$$(13)\quad Q = f'(L_2)\cdot h \cdot \int_{t=0}^{\infty} e^{-(r+b)t}dt$$
$$= f'(L_2)\cdot h/(r+b).$$

The expected present discounted value of wages paid to such a worker, W, is the sum of discounted wages paid to each of the types, weighted with $1-h$ and h, that is,

$$(14)\quad W = \hat{w}\left[(1-h)\int_{t=0}^{\infty} e^{-(r+b+q)t}dt + h\cdot\int_{t=0}^{\infty} e^{-(r+b)t}dt\right]$$
$$= \hat{w}\left(\frac{1-h}{r+b+q}+\frac{h}{r+b}\right).$$

The first-order condition for firms now implies $Q = W$, yielding, when inserting from (5) for $\hat{w}$,

$$(15)\quad f'(L_2) = \frac{r+b+qh}{(r+b+q)h} \times\left[s+e\left(1+\frac{r+b+a}{q}\cdot\frac{r+\gamma a}{r+a}\right)\right].$$

In the S-S model, the corresponding expression is, given the same macro production function,

$$(16)\quad f'(L_s) = s+e\left(1+\frac{r+b+a}{q}\right).$$

Assume now that the number of good workers is the same in both models.[3] Equations (15) and (16) permit us to readily compare the two employment rules. The following cases are of particular interest.

Case (a): When $\gamma = 1$ (no screening), w is the same for given a. Since $h<1$, $f'(L_2) > f'(L_s)$. Employment is then lower in the present case and unemployment among good workers higher. Here being fired gives no signal to future employers about worker quality. The contamination effect from having some shirking workers then dominates, making employers less willing to hire than they would be in the homogeneous labor case. From (12) and (15), this effect is stronger the larger is the share k of bad workers.

Case (b): When $\gamma = 0$ and a is finite, some type 2 workers are unemployed, while there can be no employment at all among type 1 workers. We must then have $h = 1$, and from (15)–(16), $f'(L_2) < f'(L_s)$. Employment is consequently now larger, and unemployment among type 2 workers lower, than in Shapiro-Stiglitz. In addition, $\hat{w}$ is lowered below that of the S-S model. The adverse signaling effect of being fired due to shirking, which would now imply permanent unemployment, here makes good workers abstain from shirking at a lower wage.

Case (c): When $\gamma = 0$, we may have full employment among N_2, and $a = \infty$. This occurs whenever L_2 is sufficiently small to fulfill

$$(15')\quad f'(L_2) \geq s+e(1+(r/q)).$$

Now equilibrium is of a "normal" type, and the NSC plays no role in establishing it.[4]

Since the S-S model with exogenous monitoring implies underemployment, it is thus clear that my model may imply either a

[3] This makes the maximum labor input that can be provided by the entire labor force identical for the two models. Alternatively, I could let total labor force size remain the same and increase efficiency of good workers in my model, in inverse proportion to their labor force share.

[4] Since (15′) requires $L_1 = 0$, it formally always holds only when $a \to \infty$, i.e., quitting type 2 workers would experience some unemployment but of an arbitrarily short duration on the average. This also appears more reasonable than requiring their unemployment periods to have a duration of exactly zero, i.e., $a = \infty$.

more or a less efficient outcome, depending on the precision of screening. Very imprecise screening implies greater unemployment while very precise screening leads to lower (sometimes zero) unemployment among good workers. While not very surprising, this at least gives credence to the soundness of the basic mechanisms of the S-S model when extended to the case of heterogeneous labor, given that one accepts their model for homogeneous labor.

One more fundamental difference from the S-S model should be noted in this context. When $k \to 0$, $h \to 1$ and labor is homogeneous in the limit. Yet w is lower here than in Shapiro-Stiglitz, whenever $\gamma < 1$. The reason for this is that each good worker now is more hesitant to shirk since that would mean having a positive chance of being identified as one of the few bad workers. We here have a mechanism much resembling Reinhard Selten's (1975) "trembling hand" which tends to support a socially more favorable equilibrium than that of Shapiro-Stiglitz.[5] Thus having a small number of bad workers in the economy is better than having only good workers.

III. Welfare Analysis

I now study the welfare properties of the market equilibrium in my model. Assuming a utilitarian welfare function, $s = 0$ (workers have no utility of leisure) and zero unemployment benefits,[6] the central planning problem can be written

$$(17) \qquad \max_{w, L_1, L_2} w(L_1 + L_2) - eL_2$$

subject to

$$\text{(a)} \quad w \geq e\left(1 + \frac{r+b+q}{q} \cdot \frac{r+\gamma a}{r+a}\right), \quad \text{(NSC)}$$

$$\text{(b)} \quad w(L_1 + L_2) = f(L_2)$$

(production feasibility)

$$\text{(c)} \quad L_1 = \frac{k}{1-k} \frac{\gamma(a+b)}{\gamma(a+b)+q} \cdot L_2$$

(screening).

The property (c) is found from (9) and implies that existing screening and monitoring technologies require a fixed ratio of L_1 to L_2. Note (c) can be used to eliminate L_1, making the problem virtually identical to that of Shapiro-Stiglitz. As in their model, one now finds that the social optimum generally occurs at the point in wage-employment space, where the average productivity curve, $f(L_2)/(L_1 + L_2)$, crosses the NSC curve, and the wage is made as high as possible. As in Shapiro-Stiglitz, unemployment is thus as a rule excessive at the competitive equilibrium, given exogenous monitoring and decreasing returns production functions. Profits should then be taxed away and used to subsidize wages.

In the special case when (15′) holds, implying $a = \infty$, is however the competitive equilibrium efficient in our model, since no type 2 workers are unemployed. Transfers from profits to wages then have no allocational effect, but only tend to push the wage up. Instead one could then pay part of the profits out as unemployment benefits without affecting employment, thus giving room for a certain amount of equitable redistribution at the optimal solution.

IV. Final Remarks

The model studied here represents a simple extension of the Shapiro-Stiglitz pure moral hazard model of unemployment. I also include a problem of adverse selection, in the form of a group of chronic shirkers in an

[5]A similar effect might be obtained in the S-S model, if employers there believed that shirking workers were "bad," or simply only were more reluctant to hire previous shirkers than previous quitters. In their model such employer reactions could, however, not be strictly rational, while they are in ours.

[6]As in S-S it is easy to show that with a utilitarian welfare function, it is never strictly advantageous to make unemployment benefits positive. With, for example, the other extreme of a Rawlsian maximin welfare function and permanent unemployment among type 1 workers, the social objective would however be to maximize the rate of unemployment benefits.

otherwise high-quality labor force. My analysis may give clues to how selection problems tend to interact with moral hazard in similar settings, such as product, credit and insurance markets, to tighten or relax quantity constraints that may arise under moral hazard alone. The indication is that the efficiency of screening significantly affects the outcome of this interaction.

Extensions of my model in the labor market context could also be considered, for example, to the case of reputation building by firms or the possibility of binding contracts, in which cases wages depending on tenure and performance may help solve the inefficiency problems that arise.[7] Other extensions are to more realistic types of heterogeneity (for example, continuous distributions of worker characteristics), to endogenous monitoring and screening, and to alternative ways of modeling worker types.[8] In my view, the real test of the viability of efficiency wage models of the S-S type lies in the successful application of the models to these issues.

[7]See W. Bentley MacLeod and James Malcolmson (1985) and myself (1985, 1986) for analyses of firm reputation building in similar contexts, showing that performance dependent wages always are preferred by firms. The role of bonding in implementing efficient solutions is studied in a similar context by B. Curtis Eaton and William White (1982).

[8]Realistically, a worker's type in any particular firm may be considered a random draw from a known nondegenerate distribution; the issue of matching then also becomes relevant.

REFERENCES

Eaton, B. Curtis and White, William D., "Agent Compensation and the Limits to Bonding," *Economic Inquiry*, July 1982, *20*, 330–43.

Johansen, Leif, "Some Notes on Employment and Unemployment with Heterogeneous Labor," *Nationaløkonomisk Tidsskrift*, 1982, 102–17.

MacLeod, W. Bentley and Malcolmson, James M., "Reputation and Hierarchy in Dynamic Models of Employment," Discussion Paper 8514, University of Southampton, 1985.

Selten, Reinhard, "A Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, *4*, 25–55.

Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Strand, Jon, "The Structure of Wages with Disciplining Unemployment," mimeo., University of Oslo, December 1985.

______, "Efficiency Wages, Implicit Contracts and Dual Labor Markets: A Theory of Work Habit Formation," mimeo., University of Oslo, June 1986.

Exact Consumer's Surplus and Deadweight Loss: A Correction

By ROBERT H. HAVEMAN, MARY GABAY, AND JAMES ANDREONI*

Jerry Hausman's 1981 paper in this *Review* is a clear demonstration that Hicksian-based estimates of welfare change can be measured exactly from an empirically estimated demand or supply curve. Presenting such a measure in empirical work is more precise than, and hence superior to, the common procedure of presenting estimates of Marshallian consumer's surplus (Angus Deaton, 1984).

Hausman emphasizes the import of such exact measures by a numerical illustration involving the welfare loss from the taxation of labor income. This case was cited by Hausman as one in which the deviation between the exact and Marshallian measures would be large, as the welfare change induced by the tax is a substantial proportion of the individual's base income. This note corrects an error in Hausman's calculation, and qualitatively alters the implications which can be drawn from it. It hereby demonstrates that while the Hicksian and Marshallian measures do differ in this important case, the Marshallian measure does not provide "a very poor approximation to the exact measure of welfare change" (p. 672).

Using his estimate of the labor supply function of wives, Hausman calculates the exact welfare effects of a 20 percent proportional tax on labor earnings (W_0abW_1 in Figure 1) and compares it to the Marshallian measure, W_0cbW_1. His calculated compensating variation welfare loss measure for the mean observation is $2,056, which is larger than the change in consumer's surplus, $1,315. The absolute deviation of the two measures is 45 percent. Figure 1 illustrates,

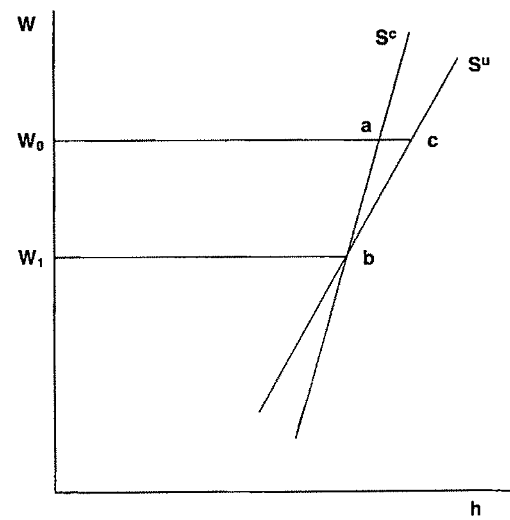


FIGURE 1

however, that the compensating variation measure should actually be smaller than consumer's surplus. Closer examination reveals that Hausman's calculated Hicksian measure is in error. Using the procedure which he correctly describes in his paper (p. 672), we find the correct value of the exact measure to be $1,247 rather than $2,056. This indicates a deviation between the two welfare measures of 5.2 percent, which corresponds to Robert Willig's formulae (1976) for the relevant income elasticity and share of income.

This 5.2 percent deviation in the welfare loss measures, it should be noted, is far less than that of the corresponding measures of the deadweight loss attributable to the 20 percent tax rate—$170.5 for the Marshallian measure vs. $231 for the Hicksian, for a deviation of 35.5 percent. Moreover, were the tax rate in the range of 30 to 40 percent (including both federal income and payroll taxes), the deviation between the exact and the Marshallian welfare loss measures would rise to 7.9 and 10.6 percent. Conversely,

*Haveman and Andreoni: Professor and Assistant Professor of Economics, respectively, University of Wisconsin, Madison, WI 53706; Gabay: Research Associate, Abt Associates, Washington, D.C. 20008.

were the actual income elasticity to equal .2 —a number consistent with a wide range of other studies of wives labor supply[1]—rather than the .6 value obtained by Hausman, the deviation of the measures of welfare change for tax rates of up to 50 percent would be less than 5 percent.

[1]These studies do not, however, account for the effect of taxes on wives labor supply.

REFERENCES

Deaton, Angus, "Demand Analysis," in Zvi Griliches and Michael Intrilligator, *Handbook of Econometrics*, Vol. III, Amsterdam: North-Holland 1984.

Hausman, Jerry A., "Exact Consumer's Surplus and Deadweight Loss," *American Economic Review*, September 1981, *71*, 662–74.

Willig, Robert D., "Consumer's Surplus Without Apology," *American Economic Review*, September 1976, *66*, 589–97.

The Relevance of Quasi Rationality in Competitive Markets: Comment

By S. KEITH BERRY*

In a recent article in this *Review* (1985), Thomas Russell and Richard Thaler demonstrate, using a Lancasterian model of consumption, that in a market with quasi-rational consumers there exist market equilibria which are not rational. The purpose of this comment is to make the following points: (a) Russell-Thaler incorrectly specified one of the demand functions in their model; (b) the necessary condition for a rational equilibrium, embodied in their Proposition 2 is incorrectly formulated; (c) the corrected condition, analogous to their Proposition 2, is both a *necessary* and *sufficient* condition; and (d) if their model is modified to assume that the weighted average of all consumer mappings from goods to characteristics is *rational*, that can result in market equilibria which are not rational.

I. The Basic Model

Assume that all individuals have the same preference over two characteristics:

$$U = C_1^{\alpha} C_2^{1-\alpha} \tag{1}$$

where C_1 and C_2 are characteristics 1 and 2, respectively. As shown by Russell and Thaler, the individual demand for characteristic 1 is

$$D_{C_1} = \alpha Y / P_{C_1} \tag{2}$$

where Y is the income of the individual (assumed the same for all individuals).

Characteristic 1 is contained only in two goods g_1 and g_2. As in Russell and Thaler, characteristic 2 is contained only in good 3.[1]

The *true* consumption relationships are given as

$$g_1 = C_1; \quad g_2 = \beta C_1. \tag{3}$$

Quasi-rational consumers believe that the relationships are

$$g_1 = C_1; \quad g_2 = \gamma C_1. \tag{4}$$

Let P_1 and P_2 be the prices of g_1 and g_2, respectively. If g_1 is bought, the price per unit of C_1 is $P_{C_1} = P_1$. If g_1 is bought, $P_{C_1} = P_2 \beta$. Note that Russell and Thaler erroneously specified $P_{C_1} = P_2/\beta$. Obviously, if 1 unit of g_2 costs P_2, with that 1 unit you get $1/\beta$ units of C_1. Thus $P_{C_1} = P_2/(1/\beta) = P_2 \beta$. Rational demands are then

$$D_1^R = ((\alpha Y/P_1), 0, (t\alpha Y/P_1)) \tag{5}$$

as $\quad P_1 < P_2\beta, P_1 > P_2\beta, P_1 = P_2\beta,$

$$D_2^R = ((\alpha Y/P_2), 0, ((1-t)\alpha Y/P_2)) \tag{6}$$

as $\quad P_1 > P_2\beta, P_1 < P_2\beta, P_1 = P_2\beta,$

where t is an arbitrary scalar $0 \le t \le 1$. The quasi-rational demands, D^{QR}, are the same expressions with γ replacing β.

Note that Russell and Thaler's specification of equation (6) is incorrect. For example, if $P_1 > P_2\beta$, the consumer purchases nothing but g_2. Since the consumer's demand for C_1 is entirely satisfied by g_2, one can express the demand for C_1 as $D_{C_1} = \alpha Y/P_{C_1} = \alpha Y/P_2\beta$. However, by equation (3) this implies that $D_2^R = \beta \cdot D_{C_1} = \alpha Y/P_2$.

If it is also assumed that there are L rational consumers, M quasi-rational consumers, and that the supply of goods 1 and 2 are fixed at $\bar{g}_1$ and $\bar{g}_2$, we are now in a position to reevaluate Russell and Thaler's Propositions 1 and 2.

*Chief Economist, Arkansas Public Service Commission, 1000 Center Street, Little Rock, AR 72201.

[1]As discussed by Russell and Thaler, we need only examine equilibrium in the g_1 and g_2 markets because of Walras' Law.

Obviously their Proposition 1 is still valid with the exception that the condition is $P_1^* = P_2^*\beta$ (where P_1^*, P_2^* is an equilibrium), and their corrected Proposition 2 is

Let $\gamma > (<)\beta$. A necessary and sufficient condition for an equilibrium to be a rational equilibrium is that $L/M \geq \bar{g}_2/\beta\bar{g}_1$ ($M/L \leq \bar{g}_2/\beta\bar{g}_1$).

To prove the sufficient condition, take the case of $\bar{g}_2/\beta\bar{g}_1 \leq L/M$ and $\gamma > \beta$. The only possible relationship between P_1 and P_2 is $P_2\beta = P_1$, a rational equilibrium. That same relationship applies to the case of $\gamma < \beta$. The proof of the necessary condition parallels that of Russell and Thaler with the corrected algebra.

There are two important implications of the corrected Proposition 2. First, in contrast with Russell and Thaler, the existence of a quasi-rational equilibrium is in no way dependent upon γ, the degree of quasi rationality. Thus, other things being equal, a small degree of quasi rationality is equivalent to a large degree of quasi rationality, when a quasi-rational equilibrium obtains.

A second implication of Proposition 2 is that if L is large enough, or if there are more than enough rational consumers, rational equilibrium will obtain. This is analogous to Russell and Thaler's conclusion regarding "too many" quasi-rational consumers causing a quasi-rational equilibrium.

II. A Model with a Rational Population Average

It could be argued that the reason Russell and Thaler's model results in the possibility of quasi-rational equilibria is that, on average, the population of consumers is quasi rational. That is, the average of the population's perceptions of the good to characteristic mappings are

$$g_1 = C_1; \quad g_2 = \frac{(L\beta + M\gamma)}{(L+M)} C_1,$$

where L consumers have mapping β and M consumers have mapping γ. When $\gamma \neq \beta$, the population average $(L\beta + M\gamma)/(L+M)$ $\neq \beta$. The rationale for this assumption is that Russell and Thaler claim that the mapping errors are in a predictable direction.

It would be interesting to see if Russell and Thaler's conclusion concerning the existence of quasi-rational equilibria is valid even if the population on average were rational. At first glance it might seem that, because individual errors cancel out in the population, the price ratio would be set as if all individuals had the population average's rational mapping, and a rational equilibrium would necessarily obtain.[2]

Assume the same model as in Section I with one exception: β is not the true mapping. Let $\beta > \gamma$ and

$$\lambda = (L\beta + M\gamma)/(L+M), \tag{7}$$

and assume that λ is the "true" mapping. If we define a rational equilibrium as that pair $\hat{P}_1, \hat{P}_2$ that would result if all $L+M$ consumers were rational ($g_2 = \lambda C_1$), then a necessary condition for an equilibrium to be rational is that $P_1^* = P_2^*\lambda$ (analogous to Russell and Thaler). Otherwise, one of the goods would not be bought at all. Furthermore,

PROPOSITION 1: $\bar{g}_2/\bar{g}_1 = M\lambda/L$ *if and only if the equilibrium is rational.*

PROOF:

If $P_1 = P_2\lambda$, it is easily shown that the consumers with mapping β purchase only good 1 and the consumers with mapping γ purchase only good 2. This implies $\bar{g}_2/\bar{g}_1 = M\lambda/L$.

Conversely, if $\bar{g}_2/\bar{g}_1 = M\lambda/L$, the only possible result is $P_2\gamma < P_1 < P_2\beta$, which implies that $\bar{g}_2/\bar{g}_1 = MP_1/LP_2$, or $P_1 = P_2\lambda$.

If the condition, $\bar{g}_2/\bar{g}_1 = M\lambda/L$, is not met then

[2]As a matter of fact, William Sharpe (1970, pp. 291–93) demonstrated in a financial model that with different investor expectations concerning return, risk, and covariance, the equilibrium that results is equivalent to that if all investors had the same *average* expectations.

PROPOSITION 2: *There exist equilibria which are not rational equilibria even when the population is rational on average.*[3]

In this model, in contrast with that from Section I, the existence of a quasi-rational equilibrium is dependent upon both quasi-rational mappings (β and γ), as well as on L, M, $\bar{g}_1$, and $\bar{g}_2$. This model also strengthens Russell and Thaler's argument that markets will not necessarily eliminate the errors. Even if the errors cancel out in the aggregate population, such that the population average is rational, a rational market equilibrium will not necessarily result.

[3]Exactly the same results are obtained in the more general case of N consumer groups, with L_i, $i=1,\dots,N$ members in each group and good to characteristic mappings for good 2 given by the equations $g_2=\beta_i C_1$. The population average and true mapping is $g_2=\lambda C_1$ where $\lambda=(\Sigma_{i=1}^N \beta_i L_i)/(\Sigma_{i=1}^N L_i)$ and Propositions 1 and 2 are exactly the same as in the two-group case.

REFERENCES

Russell, Thomas and Thaler, Richard, "The Relevance of Quasi Rationality in Competitive Markets," *American Economic Review*, December 1985, *75*, 1071–82.

Sharpe, William F., *Portfolio Theory and Capital Markets*, New York: McGraw-Hill, 1970.

The Relevance of Quasi Rationality in Competitive Markets: Reply

By THOMAS RUSSELL AND RICHARD THALER*

In our 1985 paper, we showed that if a market includes both rational and systematically biased (quasi rational) buyers, the presence of rational buyers is not enough in general to keep market prices rational. In his comment, Keith Berry (1987) has now shown that, although this conclusion is correct, the algebra leading to it is flawed. He has provided the correct derivation of the result and shown that it holds even when all agents are quasi rational, but the market is on average rational.

With the corrected algebra, a rather surprising conclusion seems to emerge. Whereas the existence of rational equilibrium depends on the *number* of rational and quasi-rational agents, it does not seem to depend on the *degree* of error of the quasi rationals. In the context of the model which we presented, this result is correct. It is not, however, robust. As we shall now show, it depends completely on an algebraic quirk.

PROPOSITION 1:

1) *Let G_i be the amount of the ith good.*

2) *Let C_i be the amount of the ith characteristic.*

3) *Let the mapping from goods to characteristics be given by a diagonal matrix H, where h_{ii} is the number of units of good i required to produce one unit of characteristic i, and where the number of goods and characteristics is n.*

4) *Let the utility function defined on characteristics be of the form*

$$U(C_1, C_2, \ldots, C_n) = C_1^{\alpha_1}, C_2^{\alpha_2}, \ldots, C_n^{1-\Sigma\alpha_i}.$$

Then, D_i, the demand for good i, is independent of h_{ii}, the entries in the diagonal matrix H.

PROOF:

By the nature of the mapping, P_{C_i} (the price of characteristic i) is given by $P_{C_i} = h_{ii}P_i$, where P_i is the price of good i. By the nature of the utility function, P_{C_i}, C_i (expenditure on C_i,) is given by $P_{C_i} \cdot C_i = \alpha_i \cdot y$, where y is income.

Thus

$$\begin{aligned} D_i &= h_{ii}C_i \\ &= h_{ii}\alpha_i y / h_{ii}P_i \\ &= \alpha_i y / P_i. \end{aligned}$$

I. A More General Model

The combination of Cobb-Douglas utility functions and diagonal mappings is thus a bad model within which to discuss the general economics of erroneous mappings. We shall, therefore, rederive Proposition 2 in the paper under less restrictive conditions.

Leaving all other assumptions of the model as is, now assume

Preferences: Let consumers' preferences be given by

$$U(C_1, C_2) = C_1^{1/2} + C_2,$$

and let income y be large enough that C_2 is bought. Normalizing prices by letting $P_{C_2} = P_3 = 1$, we have

$$D_1^R = \left(1/(2P_1)^2, 0, t/(2P_1)^2\right)$$

$$D_2^R = \left(0, \beta/(2\beta P_2)^2, (1-t)/(2P_1)^2\right)$$

as $\quad P_1 < \beta P_2, P_1 > \beta P_2, P_1 = \beta P_2.$

*Department of Economics, Santa Clara University, Santa Clara CA 95053, and Johnson Graduate School of Management, Cornell University, Ithaca NY 14853, respectively. We thank Hal Varian for helpful discussions. The usual disclaimer applies.

Quasi-rational demands for goods 1 and 2 are the same with β replaced with γ (β and γ are as defined in the original paper).

Now we have

PROPOSITION 2: *A necessary condition for an equilibrium to be a rational equilibrium is that* $P_1^* = \beta P_2^*$.

PROPOSITION 3: *Let* $\gamma < (>)\ \beta$. *Then a necessary condition for an equilibrium to be a rational equilibrium is that* $M/L \leq \gamma\bar{g}_2/\beta^2\bar{g}_1$ $(M/L \leq \beta\bar{g}_1/\bar{g}_2)$.

PROOF:

Let $\gamma < \beta$. Then at any rational price pair $P_1 = P_2\beta$ quasi rationals will not buy good 1. Clear the market in good 2 using only quasi-rational demand. Then $M\gamma/(2\gamma P_2^*) = \bar{g}_2$ so that $P_2^* = (M/4\gamma\bar{g}_2)^{1/2}$. For rationality, we require

$$P_1^* = \beta P_2^* = \beta(M/4\gamma\bar{g}_2)^{1/2}.$$

At this price total demand for good 1 is given by

$$L/\beta(M/4\gamma\bar{g}_2)^{1/2} = L\gamma\bar{g}_2/\beta^2 M.$$

We will fail to have a rational equilibrium if this is less than the total supply of good 1, that is, if $L\gamma\bar{g}_2/\beta^2 M < \bar{g}_1$, from which the inequality in the proposition follows.

Now let $\gamma > \beta$. Then the quasi rationals will not buy good 2. Repeating the argument we see that we must have rational demand for good 2,

$$L\beta/\big(2\beta(M/4\bar{g}_1)^{1/2}/\beta\big)^2 \geq \bar{g}_2$$

from which the inequality follows.

Clearly, then, in this example, when $\gamma > \beta$ the degree of error does enter the condition for rationality. On the other hand, when $\gamma < \beta$, the condition does not involve γ. Why is there this asymmetry?

In the model we have assumed that the quasi rationals make an error in evaluating only good 2. When they do not buy this good, as when $\gamma > \beta$, this error cannot affect the outcome. This result is quite general. When there are more characteristics than goods, errors in evaluating goods which are not bought cannot affect the outcome.

Note also that although the quasi-linear utility function is itself quite special, the presence of γ in the conditions for rationality does not depend in any vital way on the special nature of this function. In general when $\gamma < \beta$, Proposition 2 will look like

$$L\beta P_2(M, y, \bar{g}_2, \gamma) \geq \bar{g}_1,$$

where P_2 is the price which clears the market for good 2 when only the quasi rationals buy it. Obviously P_2 in general depends on γ (another example is the CES case), but it happens not to in the Cobb-Douglas case.

II. Conclusion

As it turns out, the model which we used to illustrate the nonexistence of rational equilibrium was not well chosen. This fact was hidden by the mathematical error. The central conclusion of the analysis is, however, correct. A rational equilibrium is not guaranteed by the presence of rational agents. The exact nature of the equilibrium, however, turns out to be quite sensitive to the assumptions made about preferences, characteristic mappings, and the number of goods and characteristics. This also may explain why other treatments of this problem (i.e., by John Haltiwanger and Michael Waldman, 1985, and by George Akerlof and Janet Yellen, 1985) though obtaining results similar to each other and to our results, also reach conclusions which in a number of respects are puzzingly different.

REFERENCES

Akerlof, George and Yellen, Janet L., "Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?," *American Economic Review*, September

1985, *75*, 708–20.

Berry, S. Keith, "The Relevance of Quasi Rationality in Competitive Markets: Comment," *American Economic Review*, June 1987, *77*, 496–98.

Haltiwanger, John and Waldman, Michael, "Rational Expectations and the Limits of Rationality: An Analysis of Heterogeneity," *American Economic Review*, June 1985, *75*, 326–40.

Russell, T. and Thaler, R. H., "The Relevance of Quasi Rationality in Competitive Markets," December 1985, *American Economic Review*, *75*, 1071–82.

Auditors' Report

February 26, 1987

Executive Committee
The American Economic Association

We have examined the balance sheets of The American Economic Association as of December 31, 1986 and 1985, and the related statements of revenues and expenses, changes in general fund and restricted fund balances and changes in financial position for the years then ended. Our examinations were made in accordance with generally accepted auditing standards and, accordingly, included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the financial statements referred to above present fairly the financial position of The American Economic Association as of December 31, 1986 and 1985, its revenues and expenses and the changes in its financial position for the years then ended, in conformity with generally accepted accounting principles which, except for the change, with which we concur, in the method of recognizing investment income as described in Note A to the financial statements, have been applied on a consistent basis.

Touche Ross and Co.
Certified Public Accountants
Nashville, Tennessee

THE AMERICAN ECONOMIC ASSOCIATION BALANCE SHEETS, DECEMBER 31, 1986 AND 1985

	1986	1985
Assets		
CASH	$ 576,067	$ 632,751
INVESTMENTS, at market (Notes A and B)	4,835,255	4,327,785
ACCOUNTS RECEIVABLE, less allowance for doubtful accounts of $590 (1986) and $4,338 (1985)	127,807	108,933
INVENTORY OF *Index of Economic Articles*, at cost	137,148	100,860
PREPAID EXPENSES	15,698	12,846
OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of $46,258 (1986) and $35,177 (1985)	65,742	61,300
	$5,757,717	**$5,244,475**
Liabilities and Fund Balances		
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	$ 368,121	$ 528,843
DEFERRED REVENUE (Note A):		
Life membership dues	41,814	44,436
Other membership dues	573,718	540,287
Subscriptions	471,646	449,263
Job Openings for Economists	20,850	19,867
	1,108,028	1,053,853
ACCRUAL FOR DIRECTORY (Note A)	138,313	68,758
FUND BALANCES:		
General	3,982,063	3,217,603
Unrecognized change in market value of investments (Notes A and C)	–	276,365
Net Worth	3,982,063	3,493,968
Restricted	161,192	99,053
Total Fund Balances	4,143,255	3,593,021
	$5,757,717	**$5,244,475**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF REVENUES AND EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1986 AND 1985

	1986	1985
REVENUES FROM DUES AND ACTIVITIES:		
Membership dues and subscriptions	$ 870,387	$ 831,233
Nonmember subscriptions	649,607	614,155
Job Openings for Economists subscriptions	32,499	30,081
Advertising	130,025	107,835
Sale of *Index of Economic Articles*	43,247	56,121
Sale of copies, republications, and handbooks	40,256	26,522
Sale of mailing list	51,127	46,346
Annual meeting	36,948	15,651
Sundry	72,514	64,191
Operating Revenues	**1,926,610**	**1,792,135**
PUBLICATION EXPENSES:		
American Economic Review	612,751	610,132
Journal of Economic Literature	767,648	779,722
Directory publication (Note A)	70,000	50,000
Job Openings for Economists	55,351	53,910
Index of Economic Articles	30,615	45,444
Journal of Economic Perspectives	75,002	–
	1,611,367	1,539,208
OPERATING AND ADMINISTRATIVE EXPENSES:		
General and administrative:		
Salaries	172,271	168,336
Rent	16,768	14,650
Other (Exhibit I)	202,862	190,503
Committee	45,286	41,728
Annual meeting	6,644	5,265
Benefit from federal income taxes (Note A)	–	(3,350)
	443,831	417,132
Operating Expenses	**2,055,198**	**1,956,340**
Operating Deficit	(128,588)	(164,205)
INVESTMENT GAINS (Note B)	248,027	449,361
REVENUES IN EXCESS OF EXPENSES	**$ 119,439**	**$ 285,156**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN GENERAL FUND BALANCE

	Total	Operations	Market Value Adjustments
Balance at January 1, 1985	**$2,830,533**	**$1,867,758**	**$ 962,775**
Add market value adjustments resulting from inflation (Note A)	101,914	–	101,914
Add revenues in excess of expenses	285,156	285,156	–
Balance at December 31, 1985	**3,217,603**	**2,152,914**	**1,064,689**
Add change in market value of investments	645,021	–	645,021
Add revenues in excess of expenses	119,439	119,439	–
Balance at December 31, 1986	**$3,982,063**	**$2,272,353**	**$1,709,710**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN RESTRICTED FUND BALANCE

	Balance at January 1	Receipts	Disburse-ments	Balance at December 31
YEAR ENDED DECEMBER 31, 1985:				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	$ 69,640	$129,000	$ 93,193	$105,447
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	–	–	5,000
The Rockefeller Foundation Grant for minority students applying for graduate work in economics	29,990	1,150	45,227	(14,087)
Sundry	5,501	100	2,908	2,693
	$110,131	**$130,250**	**$141,328**	**$ 99,053**
YEAR ENDED DECEMBER 31, 1986:				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	$105,447	$171,918	$140,836	$136,529
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	–	–	5,000
The Rockefeller Foundation Grant for minority students applying for graduate work in economics	(14,087)	125,001	95,144	15,770
Sundry	2,693	4,100	2,900	3,893
	$ 99,053	**$301,019**	**$238,880**	**$161,192**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN FINANCIAL POSITION FOR THE YEARS ENDED DECEMBER 31, 1986 AND 1985

	1986	1985
Cash, beginning of year	$632,751	$ 915,479
SOURCES OF CASH:		
Revenues in excess of expenses	119,439	285,156
Noncash charges:		
Depreciation	10,418	8,564
Directory publication (Note A)	70,000	50,000
Market value adjustments (Note A)	(248,027)	(187,415)
Cash provided by operations	(48,170)	156,305
INCREASE (DECREASE) IN CASH DUE TO CHANGES IN:		
Investments	(507,470)	(1,162,597)
Accounts receivable	(18,874)	(7,683)
Inventory of *Index of Economic Articles*	(36,288)	(9,942)
Prepaid expenses	(2,852)	9,537
Office furniture and equipment	(14,860)	(29,188)
Accounts payable and accrued liabilities	(160,722)	197,015
Deferred revenue	54,175	54,064
Accrual for directory	(445)	(192,852)
Restricted funds	62,139	(11,078)
General fund, market value adjustments	616,683	101,914
Unrecognized change in market value of investments	–	611,777
Cash, end of year	**$576,067**	**$ 632,751**

See notes to financial statements.

Notes to Financial Statements

A. Summary of Significant Accounting Policies

Investments are accounted for on a market value basis. According to the method the Association used for 1985 and earlier years, investment income included dividends, interest and inflation-adjusted capital gains or losses whether realized or not. The change in market value of corporate stocks, government obligations, bonds and commercial paper during the year, after adjusting for an inflation factor (3.3% in 1985), was recognized in income over a three-year period for corporate stocks and reflected in current income for government obligations, bonds and commercial paper. Beginning in 1986, the investment gains recognized have been modified to reflect only the Association's approximate historical average rate of return, which is currently 5%. The investment gains for 1986 represent 5% of the total cash and market value of investments at the beginning of the year. The change in market value of investments and dividends and interest earned net of investment gains recognized is recorded directly to the general fund.

The Accrual for directory results because every three to five years the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was most recently published in 1985 and distributed at no cost to the membership. In order to properly match the publishing cost of this directory with revenue from membership dues, the Association provided $70,000 in 1986 and $50,000 in 1985 for estimated publishing costs which will reduce actual directory expenses in the year of publication.

NOTES

1987 *Nominating Committee of AEA*

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Robert Eisner has appointed a Nominating Committee for 1988 consisting of Charles P. Kindleberger, Chair; Nancy S. Barrett, David I. Fand, Herbert Gintis, Ronald L. Oaxaca, Michael Rothschild, and Allen L. Sinai.

Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-Elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various officers to the Committee."

Nominations for AEA Officers: 1988

The Electoral College on March 20 chose Joseph A. Pechman as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1987. Other nominees (chosen by the 1987 Nominating Committee) are: Vice President (two to be elected), Martin Feldstein, Roy Radner, F. M. Scherer, and A. Michael Spence; for members of the Executive Committee (two to be elected), George A. Akerlof, William A. Brock, Edward M. Gramlich, and Isabel V. Sawhill.

Under a change in the bylaws as described in the *American Economic Review Proceedings*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

The fifth John Deutsch Institute Roundtable, "Rent Control: The International Experience," will take place September 1–4, 1987, at Queen's University, Kingston, Canada. The purpose of the Roundtable is to bring together acknowledged housing experts from around the world to discuss the form and effects of rent control in different jurisdictions. For full information, contact Professor Richard Arnott, Department of Economics, Queen's University, Kingston, Ontario K7L 3N6 (telephone 613+545–2294).

The Association for the Advancement of Policy, Research, and Development in the Third World announces a conference on International Development, Law, and Cooperation to be held November 18–21, 1987, in Bermuda. For full information, contact Professor Shah Mehrabi, Department of Economics, Mary Washington College, 1301 College Avenue, Fredericksburg, VA 224011 (telephone 703+899–4092/4715).

The Council for International Exchange of Scholars (CIES) announces the 1988–89 Fulbright Scholar Awards competition. Benefits include round-trip travel for the grantee and, for full academic year awards, one dependent; living costs allowance; tuition as well as book and baggage allowances. Grants are for 3–12 months in over 100 countries. Applicants must be U.S. citizens, hold the Ph.D. or comparable professional qualifications, and have university or college teaching experience. Deadlines are June 15, 1987, for Australasia, India, Latin America, and the Caribbean; September 15, 1987, for Africa, Asia, Europe, and the Middle East, and Lecturing Awards to Mexico, Venezuela, and the Caribbean; November 1, 1987, for institutional proposals for the Scholar-in-Residence Program; January 1, 1988, for Administrators' Awards in Germany, Japan, and the United Kingdom; Seminar in German Civilization; and the NATO Research Fellowships; and February 1, 1988, for Spain Research Fellowships, and France and Germany Travel-Only Awards. For more information and applications, contact CIES, Eleven Dupont Circle, NW, Washington, D.C. 20036–1257 (telephone 202+939–5401).

The John D. and Catherine T. MacArthur Foundation offers Grants for Research and Writing in International Peace and Security. The deadlines for proposals are February 15 and September 30. The grants range up to $60,000 for a single applicant, or up to $100,000 for a team project. For full information and guidelines for submissions, contact George B. Hogenson, Assistant Director, International Peace and Security Program, 140 South Dearborn Street, Suite 700, Chicago, IL 60603 (telephone 312+726–8000).

The Social Science Research Council administers a related program of dissertation and postdoctoral training and research fellowships in international security studies, with funds provided by the MacArthur Foundation. Address inquiries to the Program in International

Peace and Security Studies, SSRC, 605 Third Avenue, New York, NY 10158.

The Leonard J. Savage Award of $500 is presented annually for an outstanding doctoral dissertation in Bayesian Econometrics and Statistics. To be considered for an award, the dissertation supervisor should submit the dissertation and a letter summarizing the main results. Dissertations completed after January 1, 1977, are eligible. The closing date each year is September 1. Send submissions to Professor Arnold Zellner, Graduate School of Business, 1101 East 58th Street, Chicago, IL 60637.

The co-winners of the 1986 award are Mohan Delampady, "Testing a Precise Hypothesis Interpreting *P*-Values from a Robust Bayesian Viewpoint" (Purdue University); S. Sivaganesan, "Robust Bayesian Analysis with ε-Contaminated Classes (Purdue University), and Herman K. Van Dijk, "Posterier Analysis of Econometric Models Using Monte Carlo Integration (Erasmus University, The Netherlands).

The Academy of International Business annual meeting will be held November 12–15, 1987, at the Sheraton Plaza, Chicago, Illinois. The topic is Operating in the Global Economy: Conceptual Foundations and Practice of International Business. For full information, contact Professor Raj Aggarwal, 1987 AIB Program Chairman, Department of Finance, University of Toledo, 2801 W. Bancroft Street, Toledo, OH 43606 (telephone 419+537–2436).

Population Research and Policy Review welcomes manuscripts concerned with developing the interaction between empirical research and public policy on topics relevant to population dynamics and structure. Publication decisions are normally made within five weeks of receipt of papers. Submit two copies to the Editor: Larry D. Barnett, School of Law, Widener University, Wilmington DE 19803–0474.

The Libra Foundation seeks peace-related research: planned, in-progress, or completed but unpublished. Abstracts should be sent to Helen Raschke, c/o Libra Foundation, 3308 Kempe Street, Wichita Falls, TX 76308. The forthcoming compilation will be available summer 1987.

Call for Papers: The 1988 meeting of the History of Economics Society will be held on June 19–21, at the University of Toronto. Proposals to deliver papers (including an abstract of 250 words), organize sessions, and/or act as discussant should be sent before December 30, 1987, to President-Elect D. E. Moggridge, Department of Economics, University of Toronto, 150 St. George Street, Toronto, Canada, M5S 1A1.

Call for Papers: The International Conference of Economists, "Economic Development and the World Debt Problem," will be held September 8–11, 1987, at the University of Zagreb. The themes are: International Trade and Economic Development; International Debts and Development Strategy; International Financial System and External Debts; Rates of Exchange, Capital Market, External Liquidity, and Debt Problem; Structural Adjustment Policy and Debts; Country Risk Analysis. The official language will be English. Papers should not exceed 20 typewritten pages. The deadline is July 1. For further information, contact Professor Dr. Soumitra Sharma, Faculty of Economics, Trg J. F. Kennedya 6, 41 000 Zagreb, Yugoslavia (telephone 041+217–800 or 222–560).

Call for Papers: The Southwestern Society of Economists will meet March 2–5, 1988, at the Hyatt Regency in San Antonio, Texas, in conjunction with the annual meeting of the Southwestern Federation of Administrative Disciplines. By October 1, 1987, submit a 200-word abstract to include a cover sheet giving name, affiliation, address, telephone number, and topic area to David E. R. Gay, SSE President-Elect, Department of Economics, BA402, University of Arkansas, Fayetteville, AR 72701. Discussants and chairs should also send a cover sheet.

Call for Papers: New quarterly journals to be published beginning 1988 seek manuscripts. Detailed "Instructions for Authors" are available from individual editorial offices.

Journal of Productivity Analysis. Professor Ali Dogramaci, Editor-in-Chief, Graduate School of Management, Rutgers-The State University of New Jersey, Newark, NJ 07102.

Journal of Real Estate Finance and Economics. James B. Kau, Editor, Department of Real Estate, College of Business Administration, University of Georgia, Athens, GA 30602, and C. F. Sirmans, Editor, Department of Finance, College of Business Administration, Louisiana State University, Baton Rouge, LA 70803.

Journal of Risk and Uncertainty. Mark Machina, Editor, Department of Economics, D-008, University of California-San Diego, La Jolla, CA 92093, and W. Kip Viscusi, Department of Economics, 2003 Sheridan Road, Northwestern University, Evanston, IL 60201.

For subscription information, please write to Kluwer Academic Publishers, Zachary Rolnik, Editor, 101 Philip Drive, Norwell, MA 02061.

Call for Papers: The fourth Conference on Ukrainian Economics at the Harvard Ukrainian Research Institute, Cambridge, Massachusetts, will be held September 6–7, 1990. The conference will be devoted to an analysis of various aspects of economic conditions in the Ukraine since the early 1970's and projections for the first quarter of the twenty-first century. To present a paper, contact Professor I. S. Koropeckyj, Department of Economics, Temple University, Philadelphia, PA 19122.

Call for Papers: The annual meeting of the Midsouth Academy of Economics and Finance will be held February 17–20, 1988, at the Arlington Hotel in Hot Springs, Arkansas. Papers and comments can be included in the Proceedings issue of the *Midsouth Journal of Economics and Finance*. To participate, submit abstracts by October 15, 1987, to Professor Paul E. Merkle, MAEF General Program Chair, College of Business, Louisiana State University, 8515 Youree Drive, Shreveport, LA 71115.

Call for Papers: The Society for the Advancement of Behavior Economics (SABE) will sponsor a session at the 1987 AEA annual meeting. Papers in applied micro- and macroeconomics are invited from members and nonmembers. Submit papers, proposals, or abstracts to Ben Gilad, Department of Business Administration, Rutgers University, Newark, NJ 07102 (telephone 201 + 648–5169).

Call for Papers: The *Journal of Behavioral Economics* will devote an issue to Enretrepreneurship. Papers should be approximately 30 pages long, and combine behavioral insights with economic theories. Send submissions to the Editor, Ben Gilad, at the above address.

Call for Papers: The American Real Estate and Urban Economics Association will hold its 1987 annual meeting in conjunction with the 1987 annual AEA meeting. Anyone wishing to present a paper should submit a completed manuscript or abstract no later than June 15, 1987, to the Program Chairman: Professor James R. Follain, University of Illinois-Urbana, 434 Commerce West, 1206 S. Sixth Street, Champaign, IL 61820 (telephone 217 + 244–0951).

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to airfare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 228 East 45th Street, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Deaths

Stanley E. Boyle, Donaghey Distinguished Professor of Economics, University of Arkansas-Little Rock, March 13, 1987.

Alice C. Gorlin, professor of economics, Oakland University, March 21, 1987.

Klaus H. Hennings, professor of economics, Techische Universitat Hannover, December 27, 1986.

Promotions

Paul B. Bennett: senior research office, Federal Reserve Bank of New York, January 1, 1987.

Oscar T. Brookins: associate professor of economics, Northeastern University, July 1986.

Roger H. Goldberg: professor of economics, Ohio Northern University, September 1, 1987.

Mitchel H. Kellman: professor, City College of New York, September 1986.

David L. Roberts: assistant vice president, Foreign Exchange Function, Federal Reserve Bank of New York, January 1, 1987.

Administrative Appointments

Robert F. Allen, Air Force Institute of Technology: chairman, department of economics and finance, Creighton University, January 1, 1987.

Ke T. Hsia: chairman, department of economics, California State University-Los Angeles, September 1, 1986.

Richard E. Sylla: associate department head, economics and business, North Carolina State University, January 1, 1987.

Lester C. Thurow: dean, Sloan School of Management, Massachusetts Institute of Technology, July 1, 1987.

New Appointments

Cindy R. Alexander, University of California-Los Angeles: economist, Economic Regulatory Section, Antitrust Division, U.S. Department of Justice, February 1987.

Ralph Bristol, Office of Tax Analysis, U.S. Treasury: Whittemore School of Business, University of New Hampshire, fall 1986.

Janet Ceglowski: economist, Foreign Research Division, Federal Reserve Bank of New York, October 27, 1986.

Ahmad Faruqui: director, Competitive Market Assessment's Group, Battelle's Columbus Division, Palo Alto, California.

Jose Garcia-Medrano: vice president and senior Latin American economist, Merrill Lynch Economics, New York, February 1, 1987.

Rama Seth: economist, International Financial Markets Division, Federal Reserve Bank of New York, October 1, 1986.

Leaves for Special Appointments

Christopher Lingle, University of Natal, South Africa: visiting foreign expert, Shanghai University of Finance and Economics, PRC, February-July 1987.

Miscellaneous

Merton H. Miller, Professor of Banking and Finance, University of Chicago: Eastern Finance Association's Scholars Award, April 24, 1987.

PLEASE NOTE: The September 1987 issue of this *Review* will be the last issue to publish the Notes section. As of June 1, 1987, all items of interest should be sent to

Journal of Economic Perspectives
Woodrow Wilson School
Princeton, NJ 08544

The International Joseph A. Schumpeter Society (ISS) was founded on September 1, 1986, during its first congress in Augsburg (W. Germany) for the purpose of promoting economic research in the Schumpeterian tradition. The Society is pleased to announce the establishment of the Schumpeter Prize to be awarded every two years at the time of the Society's congress. This prize will be awarded in recognition of a recent scholarly contribution on a designated topic related to Schumpeter's work and will carry a cash award of $ 10.000 (US). The prize is endowed by the 'Wirtschaftswoche', the German economic's weekly, Düsseldorf (W. Germany).

The managing board of the Society has designated "A Study in Diffusion of Technology", as the topic for the first prize competition, where the term "technology" may be understood either narrowly (products, processes and specific technological concepts), or broadly (productive systems and organizations). Entries of major article length are preferred; a book is eligible for the prize provided that it constitutes a coherent, unitary contribution. Works submitted should be either unpublished or published subsequent to January 1, 1986.

The members of the selection committee are: Sidney G. Winter (Yale University), chairman; W. Brian Arthur (Stanford University), Bo Carlsson (Case Western Reserve University); Giovanni Dosi (DAEST, Venice); Wolfram Engels (University of Frankfurt); Horst Hanusch (University of Augsburg), ex officio Secretary General of the ISS; Carl Christian von Weizsäcker (University of Bern).

To be considered for the prize to be awarded in May 1988 at the second congress of the ISS in Siena, Italy, entries must be submitted by December 15, 1987, to the Chairman, Schumpeter Prize Committee, Yale School of Management, Box 1A, New Haven CT, 06520, USA.

Invitation: The Schumpeter Prize

Wirtschafts Woche

FAST • EASY TO USE • POWERFUL

THE GAUSS™

MATHEMATICAL AND STATISTICAL SYSTEM

Written by Lee E. Edlefsen and Samuel D. Jones

FOR IBM PC/XT/AT AND COMPATIBLES

GAUSS is a sophisticated mathematical and statistical programming package for the IBM PC and compatibles. It combines **speed, power, and ease of use** in one amazing program.

GAUSS allows you to do essentially **anything you can do with a mainframe** statistical package — and a lot more.

Personal computers are **friendly, convenient, and inexpensive**: So is **GAUSS. GAUSS** is not just a stripped-down mainframe program. **GAUSS** has been designed from the ground up to take advantage of all of the conveniences of a personal computer. After trying **GAUSS,** you may never use a mainframe again.

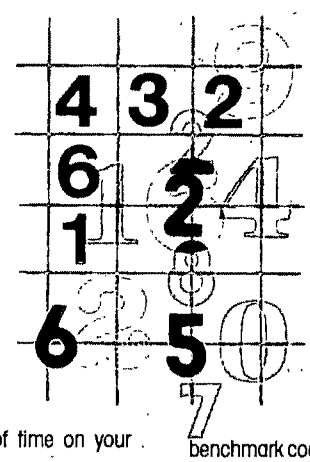
GAUSS comes with programs written in its matrix programming language that allow you to do most **statistical procedures,** including OLS, 2SLS, 3SLS, PROBIT, LOGIT, MAXIMUM LIKELIHOOD, and NON-LINEAR LEAST SQUARES.

In the current version, **GAUSS** will accept up to 90 variables in a regression. There is **no limit** on the number of observations.

GAUSS will do a regression with 10 independent variables and 800 observations in under 4 seconds — and with 50 variables and 10,000 observations in under 18 minutes. It will compute the **maximum likelihood** estimates of a binary logit model, with 10 variables and 1,000 observations, in 1-2 minutes, depending upon the number of iterations required.

GAUSS allows you to do **complicated statistical procedures** that you would never imagine trying on a mainframe. It is **easy to program** almost any routine, and **GAUSS** is so **fast** that it can do almost any job. But the nicest thing of all is that the cost of time on your personal computer is essentially zero!

GAUSS is an **excellent teaching tool.** It makes programming easy and allows students to focus on concepts and techniques.



If you can write it mathematically, you can write it in **GAUSS.** Furthermore, you can write it in **GAUSS** almost exactly the way you would write it mathematically.

GAUSS is **10-15 times faster** than other programs that use the 8087, and **15-100 times faster** than other programs that do not use the 8087.

As in APL, single statements in **GAUSS** can accomplish what might take dozens of lines in another language. However, **GAUSS** provides you with additional **powerful numerical operators and functions** — especially for statistics and the solution of linear equations — that are not found in APL. And, of course, the syntax in **GAUSS** is much more natural (for most of us) than that in APL.

GAUSS has **state-of-the-art numerical routines and random number generators.**

GAUSS is extremely accurate. It allows you to do an entire regression in 19 digit accuracy. It will compute the Longley benchmark coeficients in 5 hundredths of a second with an average of 11 correct digits! (Try that on a mainframe!)

GAUSS, with its built-in random number generators and powerful functions and operators, is an excellent tool for doing **simulations.**

GAUSS and the 8087 NUMERIC DATA PROCESSOR GIVE YOU MINICOMPUTER PERFORMANCE ON YOUR DESKTOP.

GAUSS Programming Language .... **$200.00** plus shipping/handling
(includes Applications Module 1)

GAUSS Mathematical and Statistical System .... **$300.00** plus shipping/handling
(includes Language, Applications Modules 1-11)

GAUSS is offered with a **30 day money-back guarantee.** GAUSS is not copy protected. Requires an IBM PC/XT/AT or compatible with 8087/80287, 2 DS/DD disk drives and/or hard disk, 320-640K RAM, DOS 2.1 or above. Graphics require IBM Color Graphics Adaptor or equivalent, with suitable monitor.

Call or Write

APTECH SYSTEMS, INC.

P.O. Box 6487, Kent, WA 98064
(206) 631-6679

10.11.87

THE AMERICAN ECONOMIC REVIEW

- Submit manuscripts (4 copies), 50 pages maximum, double-spaced, to:
  Orley Ashenfelter, Editor, *AER*; 209 Nassau Street, Princeton, NJ 08542–4607.

- Submission fee: $50 for members; $100 for nonmembers. Please pay with a check or money order payable in United States Dollars. Canadian and Foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Style guides will be provided upon request.

December 1987

VOLUME 77, NUMBER 5

Articles

Shorter Papers

Erratum

The following Statement of Ownership, Management, and Circulation is provided in accordance with the requirements, as contained in 39 U.S. Code 3685. The *American Economic Review* is owned, managed, and published by the American Economic Association, a nonprofit scientific organization, located at 1313 21st Avenue South, Suite 809, Nashville, Davidson County, Tennessee 37212–2786. The Editor is Professor Orley Ashenfelter, *American Economic Review*, 209 Nassau Street, Princeton, NJ 08542–4607. During the preceding 12 months the average number of copies printed for each issue was 27,600, the average paid circulation, 23,997; the average free distribution, 415; the average number of copies distributed, 24,412. Corresponding figures for the last isssue before filing: 27,500 total number copies printed; 24,397 total paid circulation; 419 free copies distributed; 24,816 total distribution.

Multicountry, Multifactor Tests of the Factor Abundance Theory

By HARRY P. BOWEN, EDWARD E. LEAMER, AND LEO SVEIKAUSKAS*

The Heckscher-Ohlin-Vanek model predicts relationships among industry input requirements, country resource supplies, and international trade in commodities. These relationships are tested using data on twelve resources, and the trade of twenty-seven countries in 1967. The Heckscher-Ohlin propositions that trade reveals gross and relative factor abundance are not supported by these data. The Heckscher-Ohlin-Vanek equations are also rejected in favor of weaker models that allow technological differences and measurement errors.

The Heckscher-Ohlin (H-O) hypothesis is most widely understood in its two-good, two-factor form: a country exports the commodity which uses intensively its relatively abundant resource. Tests of this hypothesis have been inconclusive for two reasons. First, the three pairwise comparisons required by this two × two model cannot be made unambiguously in a multifactor, multicommodity world. Most previous papers that claim to present tests of the hypothesis have used intuitive but inappropriate generalizations of the two×two model to deal with a multidimensional reality. Second, the H-O hypothesis is a relation among three separately observable phenomena: trade, factor input requirements, and factor endowments. A proper test of the hypothesis requires measurements of all three of these variables. Much prior work that claims to have tested the hypothesis has used data on only two of the three hypotheticals.

*Graduate School of Business, New York University, New York, NY 10006, and National Bureau of Economic Research; University of California-Los Angeles, Los Angeles, CA 90024; and Bureau of Labor Statistics, Washington, D.C., 20212, respectively. Bowen's work on this paper was completed while on leave as an Olin Fellow at the NBER. Sveikauskas expresses his appreciation to Rensselaer Polytechnic Institute where he conducted most of his work on this paper while on a one-year leave. The work presented here does not necessarily represent the views of the U.S. Department of Labor. This paper benefited from the comments of two referees and the participants at seminars at Harvard, MIT, and New York University. The original version of this paper was presented at the 1982 meeting of the Western Economic Association.

This paper reports conceptually correct tests of the H-O hypothesis as suggested by Edward Leamer (1980) and Leamer and Harry Bowen (1981). We use a valid multidimensional extension of the two × two model known as the Heckscher-Ohlin-Vanek (H-O-V) theorem, which equates the factors embodied in a country's net exports to the country's excess supplies of factor endowments. And we use separately measured data on trade, factor input requirements, and factor endowments to conduct the first systematic and complete evaluation of the relationships implied by the H-O-V hypothesis among these three sets of variables.

Our methods contrast sharply with traditional approaches to testing the H-O hypothesis. The classic test of the H-O hypothesis is Wassily Leontief's (1953), which compares the capital per man embodied in a million dollars worth of exports with the capital per man embodied in a million dollars worth of imports. Leamer (1980) shows this comparison does not reveal the relative abundance of capital and labor in a multifactor world. Moreover, Leontief's study uses data on trade and factor input requirements but not factor endowments and, in addition, his data are only for a single country.

A second type of purported test uses a regression of trade of many commodities on their factor input requirements for a single country (for example, Robert Baldwin, 1971; William Branson and Nicholas Monoyios, 1977; Jon Harkness, 1978, 1983; Robert Stern and Keith Maskus, 1981). If the estimated coefficient of some factor is posi-

tive, the country is inferred to be abundant in that resource. Leamer and Bowen (1981) show this also is an inappropriate inference in a multifactor world since there is no guarantee that the signs of the regression coefficients will reveal the abundance of a resource. Moreover, these studies do not use factor endowment data.[1]

A third approach used to study the sources of comparative advantage involves regression of net exports of a single commodity for many countries on measures of national factor supplies (Bowen, 1983; Hollis Chenery and Moses Syrquin, 1975; and Leamer, 1974, 1984). These papers use no measures of factor input requirements and they study the weakened hypothesis that the structure of trade can be explained by the availability of resources. This contrasts with the stricter H-O-V hypothesis studied here that factor supplies, factor input requirements, and trade fit together in a special way.

The present study computes the amount of each of twelve factors embodied in the net exports of 27 countries in 1967, using the U.S. matrix of total input requirements for 1967. The factors embodied in trade are then compared with direct measures of factor endowments to determine the extent to which the data conform to the predictions of the H-O-V theory.

We first test the traditional interpretation of the H-O hypothesis that trade reveals relative factor abundance.[2] This analysis is analogous to Leontief's attempt to determine the relative abundance of capital and labor in the United States using U.S. data alone. Our empirical results offer little support for this facet of the H-O-V model. Several types of measurement error could account for these results. Moreover, the H-O-V model implies a set of equalities, not inequalities, among the variables. We therefore extend the analysis of the H-O-V model to a regression context, and conduct a second set of tests which examine these equalities while allowing different hypotheses about consumer's preferences, technological differences, and various forms of measurement error.

Overall, our results do not support the H-O-V hypothesis of an exact relationship between factor contents and factor supplies. Support is found for the H-O-V assumption of homothetic preferences, but estimates of the parameters linking factor contents and factor supplies are found to differ significantly from their theoretical values. The data suggest that the poor performance of the H-O-V hypothesis is importantly related to measurement error in both trade and national factor supplies across countries, and the data favor a model that allows neutral differences in factor input matrices across countries.

I. Theoretical Framework

Derivation of the relationships studied here starts with the equilibrium identity expressing a country's net factor exports as the difference between factors absorbed in production and factors absorbed in consumption.

$$\mathbf{A}_i\mathbf{T}_i = \mathbf{A}_i\mathbf{Q}_i - \mathbf{A}_i\mathbf{C}_i, \tag{1}$$

where $\mathbf{A}_i = K \times N$ matrix of factor input requirements which indicate the total (direct plus indirect) amount of each of K factors needed to produce one unit of output in each of N industries,

$\mathbf{T}_i = N \times 1$ vector of net trade flows of country i,

$\mathbf{Q}_i = N \times 1$ vector of country i's final outputs,

$\mathbf{C}_i = N \times 1$ vector of country i's final consumption.

Full employment implies $\mathbf{A}_i\mathbf{Q}_i = \mathbf{E}_i$, where $\mathbf{E}_i$ is the $K \times 1$ vector of country i's factor

[1]An exception is Jon Harkness (1978, 1983), who tests the H-O-V sign and rank propositions (see below) by comparing measured factor contents with excess factor supplies that are inferred from coefficients estimated by regressing factor contents on input requirements. This analysis is suspect, however, since the estimated coefficients need not correspond either in sign or rank to a country's true excess factor supplies. See Leamer and Bowen (1981).

[2]Maskus (1985) reports conceptually correct tests of this interpretation of the H-O-V theorem for the United States using 1958 and 1972 data.

supplies. Thus, the vector of factors embodied in net trade is

$$\mathbf{A}_i\mathbf{T}_i = \mathbf{E}_i - \mathbf{A}_i\mathbf{C}_i. \tag{2}$$

This identity is transformed into a testable hypothesis by making one or more of the following three assumptions:

(A1) Assumption 1: *All individuals face the same commodity prices.*

(A2) Assumption 2: *Individuals have identical and homothetic tastes.*

(A3) Assumption 3: *All countries have the same factor input matrix*, $\mathbf{A}_i = \mathbf{A}$.

Ordinarily, the assumption of identical input matrices (A3) would be replaced by the assumption of factor price equalization and internationally identical technologies. The alternative to factor price equalization permitted here is that input requirements are technologically fixed and identical across countries, but countries have different factor prices and therefore produce different subsets of commodities.

Assumptions (A1) and (A2) imply that the consumption vector of country i is proportional to the world output vector ($\mathbf{Q}_w$), $\mathbf{C}_i = s_i\mathbf{Q}_w$, where s_i is country i's consumption share. The consumption share can be derived by premultiplying the net trade identity ($\mathbf{T}_i = \mathbf{Q}_i - s_i\mathbf{Q}_w$) by the vector of common goods prices

$$s_i = (Y_i - B_i)/Y_w, \tag{3}$$

where Y_i is GNP and B_i is the trade balance. If trade is balanced, then s_i equals country i's share of world GNP.[3]

If, in addition, the factor input matrices are identical, we can write $\mathbf{A}_i\mathbf{C}_i = s_i\mathbf{A}\mathbf{Q}_w = s_i\mathbf{E}_w$, where $\mathbf{E}_w = \Sigma_i\mathbf{E}_i$ is the $K \times 1$ vector of world factor supplies. Then, (2) can be written as

$$\mathbf{A}\mathbf{T}_i = \mathbf{E}_i - \mathbf{E}_w(Y_i - B_i)/Y_w. \tag{4}$$

Equation (4) specifies an exact relationship between factor contents and factor endowments. This relationship can be tested by measuring the net export vector $\mathbf{T}_i$, the factor input matrix $\mathbf{A}$, and the excess factor supplies $\mathbf{E}_i - s_i\mathbf{E}_w$, and computing the extent to which these data violate the equality given by (4). Such analysis requires some sensible way of measuring the distance between two matrices: the matrix with columns equal to the factor contents of trade for each country, and the matrix with columns equal to the excess factor supplies for each country. In Section II we first examine the extent to which row and column elements of these matrices conform in sign and rank without reference to any specific alternative hypotheses. In Section III we then report tests against alternatives involving nonproportional consumption, measurement errors, and differences in input matrices.

Our analysis uses data on the 367-order U.S. input-output table for 1967, and the 1967 trade and the 1966 supply of twelve resources for 27 countries.[4] The countries are those for which both occupational data and detailed trade data were available. The twelve resources are net capital stock, total labor, professional/technical workers, managerial workers, clerical workers, sales workers, service workers, agricultural workers, production workers, arable land, pastureland, and forestland.

Net capital stocks were computed as the sum of discounted real investment flows in domestic currency and converted to U.S. dollars using 1966 nominal exchange rates. Industry capital requirements (plant, equipment, and inventories) were constructed from data on U.S. industry capital stocks.

The seven labor categories are those defined at the one-digit level of International

[3] If factor prices are equalized, s_i can also be derived by premultiplying (2) by the vector of factor prices. If factor prices are unequal, (2) can still be premultiplied by the vector of factor prices prevailing in country i to obtain an expression analogous to (3), but with both internal and external factor earnings evaluated only in terms of country i's factor prices.

[4] The Data Appendix provides detailed discussion of the data.

Standard Classification of Occupations. Total labor is a country's economically active population. Input requirements for each type of labor were constructed using occupational data from the 1971 *U.S. Survey of Occupational Employment* and the 1970 *U.S. Census of Population*. Labor data are measured in numbers of people.

The three land types conform to the definitions used by the Food and Agricultural Organization. Industry land requirements were based on the U.S. input-output table; I/O sector 1 was used for pastureland, I/O sector 2 was used for arable land, and I/O sector 3 (forest and fisheries) was used for forestland. Land is measured in hectares.

Finally, data on each country's trade in 1967 were obtained at the four- and five-digit level of the Standard International Trade Classification (SITC) and concorded to input-output sectors to perform the required vector multiplications.

II. Tests of Qualitative Hypotheses

The traditional implication of the H-O theory is that factor abundance determines which commodities are exported and which are imported, in other words, the sign of net exports. In this section we report tests of the analogous qualitative implications of the H-O-V equations concerning the sign and ordering of the factor content data.

A typical kth element of (4) can be written as

$$(5)\quad \left(F_{ki}^{A}/E_{kw}\right)/(Y_i/Y_w) = \left[(E_{ki}/E_{kw})/(Y_i/Y_w)\right]-1,$$

where F_{ki} is the kth element of the factor content vector $\mathbf{F}_i=\mathbf{AT}_i$, and $F_{ki}^{A}=(F_{ki}-E_{kw}B_i/Y_w)$ is the factor content if trade were balanced. The quantity on the right-hand side of (5) is a measure of the relative abundance of resource k. If this equation is accurate, the factor content of trade can be used as an indirect measure of factor abundance. We study here two qualitative implications of (5); first, that trade reveals the abundance of resources compared with an average of all resources, and second, that trade reveals the relative abundance of resources.

The income share in (5) is an average of the resource shares weighted by world earnings: $(Y_i/Y_w)=\Sigma_k[w_kE_{ki}/\Sigma_kw_kE_{kw}]=\Sigma_k[(w_kE_{kw})(E_{ki}/E_{kw})/\Sigma_kw_kE_{kw}]$, where w_k is the world price of factor k. If equation (5) is accurate, then the sign of the net trade in factor services, corrected for the trade imbalance, will reveal the abundance of a resource, compared with other resources on the average.[5] This sign proposition is tested for each factor (country) by computing the proportion of sign matches between corresponding elements in each row (column) of the matrix of adjusted factor contents and the matrix of factor abundance ratios. In addition, Fisher's Exact Test (one-tail) is used to test the hypothesis of independence between the sign of the factor contents and of the excess factor shares against the alternative of a positive association.

Equation (5) also implies that trade reveals the relative abundance of resources when considered two at a time. If equation (5) is accurate, the adjusted net exports of country i of factor k exceed the adjusted net exports by country i of factor k', $(F_{ki}^{A}/E_{kw})/(Y_i/Y_w) > (F_{k'i}^{A}/E_{k'w})/(Y_i/Y_w)$, if and only if factor k is more abundant than factor k', $(E_{ki}/E_{kw})/(Y_i/Y_w) > (E_{k'i}/E_{k'w})/(Y_i/Y_w)$; and the adjusted net exports by country i of

[5] Other definitions of factor abundance are possible. In an earlier version of this paper, we wrote (5) without adjusting the left-hand side for the trade imbalance as $F_{ki}/s_iE_{kw}=(E_{ki}/E_{kw})-s_i$, where s_i is the consumption share $(Y_i-B_i)/Y_w$. In this form the theory can be said to imply that the sign of the net trade in factor services reveals the abundance of a factor compared with the consumption share. Equivalently, the right-hand side of this equation takes the sign of the difference between world output per factor input and the domestic consumption per factor input. This form of comparison is made by Richard Brecher and Ehsan Choudhri (1982) who point out that Leontief's findings of a positive net trade in labor services is inconsistent with the relatively high consumption per worker of the United States. Though this comparison is appropriate, we have opted here for the comparison suggested by equation (5), because it is based on a more appealing notion of factor abundance. See Kohler (1987) for a related discussion.

TABLE 1—RATIO OF ADJUSTED NET TRADE IN FACTOR TO NATIONAL ENDOWMENT

Country	Capital	Labor	Prof/Tech	Manager	Clerical	Sales	Service	Agriculture	Production	Arable	Forest	Pasture
Argentina	1.32	−0.30	−1.64	−2.60	−1.07	−0.62	−0.83	4.30	−1.46	21.24	−6.94	2.40
Australia	−3.77	−0.41	−2.95	−1.79	−1.68	0.21	−0.11	18.10	−3.65	17.15	−13.68	0.80
Austria	−2.03	3.01	2.74	5.64	2.91	3.81	3.20	3.12	2.59	−80.74	13.52	24.35
Bene-Lux	−2.36	1.81	0.88	1.82	1.90	1.36	2.39	−4.26	2.76	−364.25	−922.53	53.27
Brazil	−5.54	−0.27	−0.85	−0.49	−0.82	−0.32	−0.23	−0.04	−0.61	2.10	−0.04	−0.02
Canada	1.82	−3.49	−3.40	−2.23	−4.00	−2.73	−1.88	4.00	−6.84	12.13	6.16	2.84
Denmark	−4.89	5.82	2.37	8.70	4.25	5.08	4.51	24.56	1.21	33.57	803.73	1763.42
Finland	4.69	2.14	0.49	4.22	1.78	1.94	1.89	1.26	3.21	−24.44	30.48	434.70
France	−4.07	0.82	0.70	1.17	1.02	0.90	1.06	0.16	1.04	−21.33	−198.68	1.79
Germany	−1.05	−0.43	1.01	1.34	0.51	−1.08	−1.05	−11.86	2.07	−323.61	−377.64	−124.77
Greece	−5.50	2.93	4.48	14.95	5.37	4.49	4.68	2.20	2.02	46.92	−61.16	1.08
Hong Kong	−46.06	4.52	5.24	3.68	8.10	3.48	3.03	−14.19	6.46	−21568	−30532	−91627216
Ireland	−1.93	6.73	4.49	13.84	7.19	6.10	8.07	10.59	2.67	17.31	−129.98	72.68
Italy	−7.03	0.74	1.25	4.67	1.42	0.39	1.27	−1.73	1.87	−39.91	−431.67	−131.90
Japan	−5.47	0.10	0.44	0.48	0.33	−0.05	−0.03	−1.54	1.18	−341.42	−268.58	−1998.58
Korea	−30.51	0.61	1.53	2.85	1.81	0.76	1.73	0.27	0.85	−42.34	−29.42	1206.60
Mexico	−0.78	0.57	0.19	0.47	0.51	0.80	0.70	0.87	−0.21	12.40	5.69	0.97
Netherlands	−4.56	4.61	3.49	6.36	3.65	4.72	5.53	22.78	1.41	82.74	−719.88	330.86
Norway	−5.54	5.57	3.75	6.15	7.98	10.22	10.58	14.59	−0.06	−125.48	105.96	660.35
Philippines	−13.94	−0.10	−0.59	−0.36	−0.81	0.03	0.06	0.14	−0.81	10.47	−8.43	−17.03
Portugal	−10.31	1.92	3.92	10.85	3.75	2.83	2.72	0.63	2.49	−28.46	24.79	12.03
Spain	−6.19	3.04	4.56	13.88	4.36	4.13	3.89	2.45	2.23	−2.74	−12.00	4.92
Sweden	0.79	1.36	0.59	2.26	1.05	1.09	1.44	−0.66	2.18	−67.23	30.93	48.00
Switzerland	−5.72	3.42	4.46	11.57	3.52	5.42	4.13	−0.79	3.04	−862.95	−352.36	−12.18
UK	−12.86	0.63	1.77	2.04	1.37	1.30	1.32	−18.57	1.11	−313.42	−2573.99	−91.89
US	0.08	−0.25	0.23	−0.11	−0.19	−1.10	−0.68	1.54	−0.34	19.45	−23.82	−1.63
Yugoslavia	−3.15	0.68	0.39	1.59	1.12	2.05	1.15	0.46	0.76	−0.08	2.81	14.24

Note: Numbers in percent. Factor content data are for 1967; endowment data are for 1966.

factor k exceeds the adjusted net exports by country i' of factor k, $(F^A_{ki}/E_{kw})/(Y_i/Y_w) > (F^A_{ki'}/E_{kw})/(Y_{i'}/Y_w)$, if and only if country i is more abundant in factor k than country i', $(E_{ki}/E_{kw})/(Y_i/Y_w) > (E_{ki'}/E_{kw})/(Y_{i'}/Y_w)$. More generally, for each country and factor, the ranking of adjusted net factor exports F^A_{ki}/E_{kw} should conform to the ranking of factors by their abundance. This rank proposition is tested for each country (factor) by computing the Kendall rank correlation between corresponding columns (rows) of the matrix of adjusted factor content and the matrix of factor abundance ratios. In addition, we compute the proportion of correct rankings when the corresponding elements of the columns (rows) of the two matrices are compared two at a time.[6]

Table 1 summarizes the factor content data by listing for each country the ratio of adjusted net exports of each factor in 1967 to the endowment of the corresponding factor in 1966, $100 \times F^A_{ki}/E_{ki}$. According to these data, the United States exports .08 percent of the services of its capital stock, .23 percent of the services of its professional/technical workers but imports labor services amounting to .25 percent of the services of its labor force. Thus, among these resources, U.S. trade reveals the United States to be most abundant in professional and technical workers, capital, and then labor. Among all resources, however, the United States is revealed most abundant in arable land, followed by agricultural workers.

Leamer (1980) computed these factor content ratios using Leontief's 1947 data and found that U.S. trade revealed the United

[6]Subsequent tests of the rank and sign propositions based on the proportion of "successes" do not refer to any specific alternative hypothesis and thus leaves unclear the choice of significance level. Without knowing the proportion of successes expected under a specific alternative hypothesis, judging the relative performance of the H-O-V model is largely impressionistic. The absence of alternative hypotheses when testing the sign and rank propositions is, in large part, the motivation for our subsequent tests of the H-O-V equations in a regression framework.

States to be abundant in capital compared to labor, thus reversing Leontief's paradoxical finding. Likewise, no "Leontief paradox" is evident in Table 1 since the United States exports capital services but imports labor services, and this ordering conforms to the ordering of the U.S. shares of world capital (41 percent) and world labor (22 percent). This result, and others like it, would lead us to accept the H-O theorem on the basis of a rank test.

Although a rank test supports the two-factor version of the H-O theorem for the United States, a contrary finding is that while the United States is a net exporter of capital services, the U.S. share of world income (47 percent) exceeds its share of world capital, which implies that there is a measured scarcity of capital in the United States. This result, and others like it, would lead us to reject the H-O theorem using a sign test.

Some obvious anomalies in Table 1 are that, after adjusting for trade imbalances, Denmark, Finland, Korea, the Netherlands, and Norway export more than 100 percent of the services of their pastureland. These anomalies probably reflect difficulties in applying U.S. input-output coefficients to other countries. For example, Denmark is a substantial exporter of agricultural products and U.S. input coefficients apparently overstate the amount of pastureland used per unit of output in Denmark. The analysis conducted in Section III will formally test the assumption of identical input coefficients, but it is clear from the anomalies in Table 1 that assumption (A3) is not entirely accurate.[7]

Formal tests of the conformity of the adjusted net factor export data (F^A_{ki}/E_{kw}) with the factor abundance data $[(E_{ki}/E_{kw})/(Y_i/Y_w)-1]$ are reported in Tables 2 and 3. The first column of Table 2 lists the proportion of sign matches between adjusted net factor exports and the abundance ratios for each factor. The first column of Table 3 lists comparable percentages for each country. For example, the sign of adjusted net capital exports and of excess capital shares matched in 52 percent of the countries.

TABLE 2—SIGN AND RANK TESTS, FACTOR BY FACTOR

Factor	Sign Test[a]	Rank Tests[b]	
Capital	.52	0.140	.45
Labor	.67	0.185	.46
Prof/Tech	.78	0.123	.33
Managerial	.22	−0.254	.34
Clerical	.59	0.134	.48
Sales	.67	0.225	.47
Service	.67	0.282[c]	.44
Agricultural	.63	0.202	.47
Production	.70	0.345[c]	.48
Arable	.70	0.561[c]	.73
Pasture	.52	0.197	.61
Forest	.70	0.356[c]	.65

[a] Proportion of 27 countries for which the sign of net trade in factor matched the sign of the corresponding factor abundance.

[b] The first column is the Kendall rank correlation among 27 countries; the second column is the proportion of correct rankings out of 351 possible pairwise comparisons.

[c] Statistically significant at 5 percent level.

In general, the proposition of conformity in sign between factor contents and excess factor shares receives relatively little support when tested for each factor (Table 2). Although the proportion of sign matches exceeds 50 percent for eleven resources, the proportion of sign matches is 70 percent or greater for only four of the twelve factors with the highest proportion of sign matches for professional and technical workers (78 percent). Moreover, using Fisher's Exact Test, the hypothesis of independence between the sign of the factor contents and of the excess factor shares can be rejected (results not shown) at the 95 percent level for only one resource—arable land.

Similar results are obtained when the sign proposition is tested for each country (Table 3). The proportion of sign matches exceeds 50 percent for 18 countries, and exceeds 90 percent for five countries (Greece, Hong Kong, Ireland, Mexico, and the UK). However, the proportion of sign matches is below 70 percent for 19 of the 27 countries. In addition, the hypothesis of independence

[7] These anomalous data values may also reflect errors of measurement in either the factor contents or endowments. In particular, Denmark and Norway probably export more than 100 percent of their forestland because these countries export fish and fish products, and fisheries are included in the input-output coefficients for forestland.

TABLE 3—SIGN AND RANK TESTS, COUNTRY BY COUNTRY

Country	Sign Tests[a]	Rank Tests[b]	
Argentina	.33	0.164	.58
Australia	.33	−0.127	.44
Austria	.67	0.091	.56
Belgium-Luxembourg	.50	0.273	.64
Brazil	.17	0.673[c]	.86
Canada	.75	0.236	.64
Denmark	.42	−0.418	.29
Finland	.67	0.164	.60
France	.25	0.418	.71
Germany	.67	0.527[c]	.76
Greece	.92	0.564[c]	.80
Hong Kong	1.00	0.745[c]	.89
Ireland	.92	0.491[c]	.76
Italy	.58	0.345	.69
Japan	.67	0.382	.71
Korea	.75	0.345	.69
Mexico	.92	0.673[c]	.86
Netherlands	.58	−0.236	.38
Norway	.25	−0.236	.38
Philippines	.50	0.527[c]	.78
Portugal	.67	0.091	.56
Spain	.67	0.200	.62
Sweden	.42	0.200	.62
Switzerland	.67	0.382	.69
United Kingdom	.92	0.527[c]	.78
United States	.58	0.309	.67
Yugoslavia	.83	−0.055	.49

[a]Proportion of 12 factors for which the sign of net trade in factor matched the sign of the corresponding excess supply of factor.

[b]The first column is the Kendall rank correlation among 11 factors (total labor excluded); the second column is the proportion of correct rankings out of 55 possible pairwise comparisons.

[c]Statistically significant at the 5 percent level.

between the classification of signs is rejected (95 percent level) for only four countries: Greece, Ireland, Hong Kong, and the United Kingdom.[8] Finally, for the entire sample, the proportion of sign matches out of a possible 324 is only 61 percent.

The sign proposition deals with the abundance of a resource compared with a value-weighted average of other resources (that is, Y_i/Y_w), but we can also compare resources two at a time. For example, the data in Table 1 indicate the United States is more abundant in capital than labor while the U.S. resource share data (not shown) also indicate an abundance in capital compared to labor. The many possible pairwise comparisons are summarized by the rank proposition, which states that the order of adjusted factor contents and the order of the resource abundance ratios conform.

Two formal measures of the conformity between the factor content and factor abundance rankings are shown in Tables 2 and 3. The second column in these tables shows the Kendall rank correlation between the rankings while the third column shows the proportion of correct orderings when the comparisons are made two at a time.[9] For example, the results for capital in Table 2 indicate that we cannot reject (5 percent level) the hypothesis of a zero-rank correlation and that the proportion of correct orderings when the ranking between the net exports of capital services and the capital abundance ratios is compared for all pairs of countries is 45 percent.

In general, the rank proposition receives little support when tested for each factor (Table 2). The hypothesis of a zero-rank correlation is rejected (95 percent level) for only four resources (service workers, production workers, arable land, and forestland) and one of the correlations (managerial workers) is of the wrong sign. Little support is also found for the rank proposition when the comparisons are made among all possible pairs of countries. Specifically, the proportion of correct orderings exceeds 50 percent only for the three land variables.

The rank proposition also receives little support when tested country by country (Table 3). The hypothesis of a zero-rank correlation is rejected for only eight of the 27 correlations (95 percent level) and five of the correlations are of the wrong sign. Somewhat greater support is found for the rank proposition when pairwise comparisons are considered: for 22 of the 27 countries, the proportion of correct orderings exceeds 50 percent. That the rank proposition re-

[8]No variation was observed in the sign of factor abundance for Yugoslavia (each was positive).

[9]These proportions are interpreted as the probability, for a given factor (country), that the ranking of factor contents will match the ranking of factor abundance for a randomly selected pair of countries (factors).

ceives relatively more support when tested country by country suggests that something is affecting all the data similarly, since adding a number that is constant within a country would not affect the country rank test results but would alter the other three tests. A possible source of this kind of problem would be differences in factor input matrices across countries.

Overall, the results for the sign and rank propositions offer little support for the H-O-V model. However, the tests of these propositions do not refer to specific alternative hypotheses and may cast doubt on the H-O-V hypothesis for a variety of reasons, including nonproportional consumption, various kinds of measurement error, and differences in factor input matrices. These alternatives can be studied by regressions of factor contents on endowments as described below.

III. Tests of the H-O-V Equations

The tradition since Leontief's study has been to examine only propositions concerning factor rankings. But as shown in Section I, the H-O-V model actually implies an equality between factor contents and resource supplies. A study of this system of equations has the advantage that it allows explicit consideration of alternative hypotheses—a practice that has generally been absent in empirical tests of trade theory. Here we consider three reasons why the H-O-V equations may be inexact: nonproportional consumption, measurement errors, and technological differences.

A. *Alternative Hypotheses*

We first consider an alternative to the assumption of proportional consumption (A2). The general hypothesis of nonidentical, nonhomothetic tastes cannot be allowed since then trade, which is the difference between production and consumption, would be completely indeterminate.[10] Instead, we study a specific alternative to assumption A2:

($\tilde{\text{A}}$2) *All individuals have identical preferences with linear Engel curves; within each country, income is equally distributed.*

The modification of (4) implied by ($\tilde{\text{A}}$2) is derived by noting that ($\tilde{\text{A}}$2) implies that per capita consumption is a linear function of per capita income. Therefore, we can write country i's total consumption of commodity j (C_{ij}) as a linear function of its population L_i and its total expenditure $(Y_i - B_i)$:[11]

$$(7)\quad C_{ij} = \lambda_j L_i + \psi_j\big((Y_i - B_i) - L_i y^0\big),$$

where λ_j = per capita "autonomous" consumption of commodity j,

ψ_j = marginal budget shares, $\sum_j \psi_j = 1$,

$y^0 = \sum_j \lambda_j$.

Summing (7) over i gives the marginal budget shares ψ_j:

$$(8)\quad \psi_j = (Q_{wj} - \lambda_j L_w)/(Y_w - L_w y^0),$$

where L_w is world population. Inserting (8) into (7) and premultiplying by the kth row of $\mathbf{A}(\mathbf{a}_k)$, the amount of factor k absorbed in consumption $\mathbf{a}_k\mathbf{C}_i$ is

$$(9)\quad \mathbf{a}_k\mathbf{C}_i = (\varphi_k - \beta_k y^0)L_i + \beta_k Y_i,$$

where

$$\varphi_k = \sum_j a_{kj}\lambda_j,$$

$$\beta_k = \Big(\sum_j a_{kj}Q_{wj} - \sum_j a_{kj}\lambda_j L_w\Big) \Big/ (Y_w - L_w y^0),$$

$$\beta_k = (E_{kw} - \varphi_k L_w)/(Y_w - L_w y^0).$$

[10] In the sense that complete information on each country's preferences would be required to determine trade.

[11] Equation (7) is based on the Linear Expenditure System.

Equation (9) implies that equation (4) can be written

$$(10)\qquad \mathbf{F}_i = \mathbf{E}_i - \theta L_i - \beta(Y_i - B_i),$$

where θ and β are $K \times 1$ vectors with elements $\theta_k = (\varphi_k - \beta_k y^0)$ and β_k, respectively. Given (10), assumption (A2) amounts to restricting $\theta = 0$ and $\beta_k = E_{kw}/Y_w$.

Next we allow for measurement errors. We assume measurement of net trade differs from its true value by a constant plus a random error

$$(\tilde{\text{M}}1)\qquad \mathbf{T}_i^m = \omega + \mathbf{T}_i + \mathbf{T}_i^e,$$

where the vector $\mathbf{T}_i^m$ is the measured value of the vector $\mathbf{T}_i$, ω is an $N \times 1$ vector of constants, and $\mathbf{T}_i^e$ is the error vector. The null hypothesis is that there is no measurement error bias

$$(\text{M}1)\qquad \omega = 0.$$

Assumption ($\tilde{\text{M}}$1) implies the factor content vector is also measured with error:

$$(11)\qquad \mathbf{F}_i^m = \mathbf{A}\mathbf{T}_i^m = \mathbf{A}\omega + \mathbf{A}\mathbf{T}_i + \mathbf{A}\mathbf{T}_i^e$$
$$= \alpha + \mathbf{F}_i + \mathbf{F}_i^e,$$

where $\mathbf{F}_i^m$ is the measured value of $\mathbf{F}_i$, $\alpha = \mathbf{A}\omega$ is a $K \times 1$ vector of unknown constants, and $\mathbf{F}_i^e$ is the error vector with covariance matrix that is assumed diagonal for convenience.

The measurements of the endowments are also assumed to be imperfect but in a different way:

$$(\tilde{\text{M}}2)\qquad E_{ki} = \gamma_k E_{ki}^m,$$

where E_{ki}^m is the measured value, E_{ki} the true value, and γ_k is a positive error multiplier. The null hypothesis of no measurement errors is

$$(\text{M}2)\qquad \gamma_k = 1 \text{ for all } k.$$

The form of the measurement error contained in ($\tilde{\text{M}}$2) is also chosen for convenience since random-measurement errors in more than one variable would force us into consideration of an "errors-in-variables" model, which entails regressions in more than one direction. With our assumptions, factor contents are always the dependent variable.

A third source of measurement error we consider is the incomplete coverage of countries. World endowments and world GNP are estimated here by summing across the sample of countries. The resulting underestimates of the world totals would not affect our analysis if excluded countries had total endowments proportional to the sample totals. As an alternative to this assumption, we can assume that the calculated totals contain no information about world totals. This latter assumption can be stated formally as

$$(\tilde{\text{M}}3)\qquad E_{kw} = \sigma_{kS} E_{kS}.$$
$$Y_w = \phi_S Y_S.$$

The subscript S refers to the subset of countries in the sample; σ_s is a set of unknown positive elements; and ϕ_s is an unknown positive scalar. The null hypothesis is

$$(\text{M}3)\qquad \sigma_{kS} = 1 \quad \text{for all } k \text{ and } \phi_S = 1.$$

Combining the assumption of nonproportional consumption ($\tilde{\text{A}}$2) with the measurement error assumptions $\tilde{\text{M}}$1–$\tilde{\text{M}}$3, the expression for country i's net trade in factor k becomes

$$(12)\qquad F_{ki} = \alpha_k + \gamma_k E_{ki} - \theta_k L_i - \beta_k(Y_i - B_i) + F_{ik}^e,$$

where the superscript "m" is suppressed for notational convenience.

The third source of alternative hypotheses is technological differences. The alternative to the assumption of identical input matrices (A3) that we consider is the assumption that input matrices differ by a proportional constant. This amounts to assuming neutral differences in technology across countries.[12]

[12] The specification of neutral technological differences was chosen because of its tractability in estimation.

TABLE 4—ALTERNATIVE ASSUMPTIONS AND PARAMETER RESTRICTIONS

	Assumptions[a]						Parameter Restrictions				
Hypothesis	A1	A2	A3	M1	M2	M3	θ_k	δ_i	α_k	γ_k	β_k
HG	*										
H1	*	*	*			*	0	1			E_{ks}/Y_s
H2	*	*		*	*	*	0		0	1	E_{ks}/Y_s
H3	*	*				*	0				E_{ks}/Y_s
H4	*	*	*	*	*		0	1	0	1	
H5	*	*	*				0	1			
H6	*	*		*	*		0		0	1	
H7	*	*					0				
H8	*		*	*	*			1	0	1	
H9	*		*					1			
H10	*			*	*				0	1	

[a]Absence of an asterisk indicates selection of the alternative $\tilde{A}_i$ or $\tilde{M}_i$. Each parameter restriction is listed in the same order as the corresponding assumptions A2–M3.

Definitions: A1 = identical commodity prices; A2 = identical and homothetic tastes; A3 = identical input intensities; M1 = unbiased measurement of factor contents; M2 = perfect measurement of endowments; and M3 = complete coverage of countries.

Since we calculate factor contents using the U.S. input matrix, the proportional difference in input matrices is measured relative to the U.S. input matrix. This assumption can be written

$$(\tilde{\text{A}}3) \qquad \mathbf{A}_{us} = \delta_i \mathbf{A}_i,$$

where $\delta_i > 0$ and $\delta_{us} = 1$.

Assumption ($\tilde{\text{A}}$3) implies that the parameters θ_k and β_k, and the values F_{ki}, are now θ_k/δ_i, β_k/δ_i and F_{ki}^{us}/δ_i, respectively, where F_{ki}^{us} is country i's net trade in factor k computed using the U.S. input matrix. Substituting these new values into (12) gives

$$(13) \quad F_{ki}^{us}/\delta_i = (\alpha_k/\delta_i) + \gamma_k E_{ki} - (\theta_k/\delta_i) L_i - (\beta_k/\delta_i)(Y_i - B_i) + F_{ki}^e/\delta_i.$$

The γ_k are not scaled by δ_i since the endowments are measured independent of the input matrix. Multiplication of (13) by δ_i yields the bilinear form

$$(14) \quad F_{ki}^{us} = \alpha_k + (\delta_i \gamma_k) E_{ki} - \theta_k L_i - \beta_k (Y_i - B_i) + F_{ki}^e.$$

Equation (14) identifies our most general model,[13] which we estimate using an iterative maximum likelihood procedure discussed below.

In addition to the general hypothesis contained in (14) (hereafter denoted HG), we consider ten alternative hypotheses H1–H10 selected from the set of possibilities corresponding to different choices from the list of assumptions about the theory and the nature of measurement errors. Table 4 states each alternative in terms of the restrictions it imposes on the parameters of (14).

Data limitations prevented us from considering more general specifications, such as allowing input requirements to differ across industries and countries.

[13]This specification was selected after testing it against the more general specification

$$F_{ki}^{us} = \pi_k + \delta_i [\alpha_k + \gamma_k E_{ki}] - \theta_k L_i - \beta_k (Y_i - B_i) + F_{ki}^e,$$

where π_k is an unknown constant.

Hypotheses HG–H10 each maintain the assumption of common goods prices (A1). Hypotheses H1–H7 further maintain the assumption of proportional consumption while allowing tests of the assumptions of identical input matrices (A3), measurement error in trade and the endowments, and incomplete coverage of countries. The hypotheses of special interest are: H4, which leaves only β_k unrestricted and corresponds to the H-O-V hypothesis that the parameter-linking factor contents and national factor supplies is unity; H3, which maintains the assumptions of proportional consumption (A2) and complete coverage of countries (M3); H9, which maintains only the assumption of identical technologies (A3); and H10, which maintains the hypothesis that both trade and the endowments are measured without error (M1 and M2).

B. *Measuring Performance and Estimation Issues*

Given estimates of the unrestricted parameters in (14) under each hypothesis, a method is required to determine the overall performance of each alternative. One possibility is to form indexes based on the maximized value of the likelihood function associated with (14)

$$L = (\mathrm{ESS})^{-(NK/2)}, \tag{15}$$

where ESS is the error sum-of-squares (summed over countries and factors) and NK is the total number of observations. Values of L, like an R^2, necessarily increase as the number of parameters increases and some form of degrees of freedom correction is required. We adopt the asymptotic Bayes' formula proposed in the context of regression by Leamer (1978, p. 113) and more generally by G. Schwarz (1978):

$$L^* = L(NK)^{-(p/2)}, \tag{16}$$

where p is the number of parameters estimated under a given hypothesis. Given an alternative hypothesis j and a null hypothesis i, we form the ratio

$$\Lambda = \mathbf{L}_j^*/\mathbf{L}_i^* \tag{17}$$
$$= (\mathrm{ESS}_i/\mathrm{ESS}_j)^{(NK/2)}(NK)^{(p_i - p_j)/2}.$$

The evidence is then said to favor the alternative if $\Lambda > 1$. If the parameter values associated with each hypothesis are considered equally likely a priori, then Λ is interpreted as the posterior odds in favor of the alternative.

The variances of the residuals in equation (14) are assumed to be different for different factors. Processing of the data would be relatively easy if these variances were all equal. For example, if the endowments were measured without error ($\gamma_k = 1$), then equation (14) could be estimated by ordinary least squares with dummy variables. But the assumption of equal variances makes little sense unless the data are scaled in comparable units. To achieve comparability, we scale all the data by the sample "world" endowment levels E_{kS}. Furthermore, to eliminate heteroscedasticity associated with country size, we also divide by the adjusted GNP: $Y_i - B_i$. After these adjustments, equation (14) becomes

$$F_{ki}^{us}S_{ki} = \alpha_k S_{ki} + \gamma_k\delta_i(E_{ki}S_{ki}) - \theta_k(L_iS_{ki}) - \beta_k E_{kS}^{-1} + F_{ki}^{e*}, \tag{18}$$

where $S_{ki} = [(Y_i - B_i)E_{ks}]^{-1}$. The errors F_{ki}^{e*} are assumed to be normally distributed with mean zero and variance σ^2.

Given observations on factor contents, resource supplies, and population, the parameters in (18) are estimated using an iterative procedure, which solves the set of first-order conditions for maximizing the likelihood function (15). Given estimates δ_i^0 ($=1$ initially), estimates α_k^0, γ_k^0, θ_k^0, and β_k^0 are obtained from a regression equation for each factor as

$$F_{ki}^{us}S_{ki} = \alpha_k S_{ki} + \gamma_k(\delta_i^0 E_{ki}S_{ki}) - \theta_k(L_iS_{ki}) - \beta_k E_{ks}^{-1} + F_{ki}^{e*}. \tag{19}$$

The estimates α_k^0, γ_k^0, θ_k^0, and β_k^0 are then

used to obtain new estimates δ_i^0 from a regression equation for each country as

$$W_{ki} = \delta_i(\gamma_k E_{ki}/E_{kS}), \tag{20}$$

where $W_{ki} = F_{ki}^{us}S_{ki} - \alpha_k^0 S_{ki} - \theta_k^0(L_i S_{ki}) - \beta_k^0 E_{ks}^{-1}$. Prior to using the new estimates of δ_i obtained from (20) to re-estimate (19), each estimate of δ_i is divided by the estimated value for the United States to maintain the restriction that $\delta_{us} = 1$. The process of iteratively estimating (19) and (20) continues until the value of (15) converges.

The above two-step procedure is used to estimate the parameters in (18) under hypotheses HG, H3, and H7 since each involves the specification that $\gamma_k \neq 1$ and $\delta_i \neq 1$. Estimates of the unrestricted parameters under hypotheses H1, H5, and H9 are estimated using OLS while the parameters under hypotheses H2, H4, H6, H8, and H10, which restrict $\gamma_k = 1$, are estimated using a dummy variables model applied to the data set pooled across countries and factors, and imposing the restriction $\delta_{us} = 1$.

C. *Analysis*

Table 5 reports information on the performance of each hypothesis. The second column of Table 5 indicates the value of the error sum-of-squares (ESS) for each hypothesis. The ESS is of course smallest for the least-restricted model (HG), although hypotheses H3 and H7 do almost as well. The corresponding log-likelihood values are reported in the next column.

Conventional hypothesis testing would compare the difference between these log-likelihood values with χ^2 values at arbitrarily selected levels of significance. For example, the χ^2 statistic for testing H3 against the unrestricted hypothesis is 58.6 ($= 2[-41.1 - (-70.4)]$), which would be compared against a number like 33.92, the upper 5 percent of a χ^2 random variable with 22 degrees of freedom (the number of restrictions). The suggested conclusion is then that the restrictions embodied in hypothesis H3 can be rejected in comparison with the unrestricted model HG. But this kind of treatment inadequately deals with the power of the test, which is inappropriately allowed to grow with the sample size while the significance level is held fixed. This emphasis on power leads to tests that avoid type II errors merely by rejecting the alternative hypothesis and it creates a serious tendency to reject restrictions as the sample size grows. This problem is alleviated here through the use of the asymptotic Bayes' factor (17), which has a certain arbitrariness in construction, but nonetheless has the effect of lowering the significance level as the sample size grows and thus maintaining some reasonable relationship between the significance level and the power.

The fifth column of Table 5 reports the log-likelihood values adjusted for the dimensionality of the parameter space according to (16). A constant has been added to these numbers so that they are all nonnegative. The corresponding Bayes' factors (or odds ratios) are reported in the last column. The clear winner is hypothesis H3, which allows neutral differences in factor input matrices, biased measurements of factor contents, and multiplicative errors in the endowments,[14] but maintains the assumptions of identical homothetic tastes and complete coverage of countries. Second best (though far behind) is hypothesis H7, which weakens H3 by allowing for incomplete coverage of countries. The third-best hypothesis is HG, the unrestricted model. The other hypotheses are essentially "impossible," given the data evidence. Such extreme values for the Bayes' factors are not uncommon, and should

[14] To examine the potential extent of measurement error in the endowments, we compared measured U.S. endowments with the amount of each factor absorbed directly and indirectly in producing the 1967 vector of U.S. final demand in both manufacturing and services (a total of 354 sectors). The ratio of the amount absorbed in production to the endowment for each factor was: capital 2.1; total labor, .88; prof/tech, .62; managerial, .45; clerical, .92; sales, 1.41; service, .68; agricultural, .98; production, .99. The discrepancy for capital likely occurs because the depreciation rates used in computing industry capital stocks were typically lower than the rate used to compute national capital stocks. The discrepancy for managerial workers likely reflects the exclusion of government employees in calculating industry input requirements.

TABLE 5—PERFORMANCE STATISTICS FOR ALTERNATIVE HYPOTHESES

Hypothesis	ESS[a]	$\ln(L)$	Number of Parameters	Adjusted[b] $\ln(L)$	Odds of Hypothesis[c] Relative to H3
HG	1.32	−41.1	71	808.1	3.15E-15
H1	6.63	−280.9	22	707.8	nil
H2	14.56	−397.7	27	576.8	nil
H3	1.61	−70.4	49	841.5	1.0
H4	961.80	−1020.0	11	0.0	nil
H5	6.35	−274.6	33	682.8	nil
H6	11.85	−367.2	38	576.0	nil
H7	1.51	−60.9	60	819.6	32.20E-10
H8	492.39	−920.6	22	68.1	nil
H9	6.25	−272.1	44	653.9	nil
H10	11.58	−363.7	49	548.1	nil

[a] In millions.
[b] Adjusted $\ln(L) = \ln(L) - (p/2)\ln(297) + 1051$, where p = number of parameters and 1051 is the value of equation (16) under hypothesis H4.
[c] Odds = exp[adjusted $\ln(L) - 841.5$]. "Nil" entries indicate a value less than 10^{-50}.

probably be viewed with suspicion since they depend on a number of assumptions, normality being a potentially important example.

Although hypothesis H3 is favored, it does not lead to sensible estimates of many of the parameters. Table 6 reports estimates of the technological differences δ_i. The hypothesis that the technology is the same as that of the United States, $\delta_i = 1$, can be rejected for all but three countries (Australia, Canada, and Mexico),[15] but most of the estimates are wildly different from one, and eight take on implausible negative values. Furthermore, 15 countries have estimated δ's in excess of one, indicating that their factors are more productive than those of the United States.

It is possible that these peculiar estimates are due to one or more "rogue" observations. Table 1 indicates that eight countries with negative estimates all have large imports of the services of one or more of the land factors: arable land, forestland, and pastureland, and accounting for these extreme values

[15] The Bayes' criterion in equation (17) implies a critical t-value of 2.19. The critical value is computed as $[(T-k)(T^{1/T}-1)]^{1/2}$, where T is the number of observations (297) and k is the number of parameters (49). See Leamer (1978, p. 114) for discussion.

TABLE 6—H-O-V REGRESSIONS AND COUNTRY COEFFICIENTS UNDER HYPOTHESIS H3

Country	δ_i^a	Standard Error	t-Statistics[b]
Argentina	1.5769	0.0941	6.129
Australia	1.1315	0.0751	1.751
Austria	3.9479	0.8720	3.380
Belgium-Luxembourg	−7.1774	2.7668	−2.955
Brazil	0.1327	0.0474	−18.281
Canada	0.9431	0.1225	−0.463
Denmark	7.2536	0.6196	10.092
Finland	4.4885	0.2966	11.758
France	−0.7803	0.7591	−2.345
Germany	−16.9248	2.0573	−8.712
Greece	6.1582	0.2809	18.357
Hong Kong	−174.4016	24.7673	−7.081
Ireland	13.4523	0.4147	30.024
Italy	−1.5930	0.7419	−3.494
Japan	−21.3424	2.2211	−10.059
Korea	3.0928	0.2646	7.906
Mexico	1.1999	0.1121	1.782
Netherlands	18.5644	3.2888	5.340
Norway	13.0655	0.8802	13.706
Philippines	2.2965	0.1057	12.258
Portugal	1.9940	0.1640	6.060
Spain	0.3709	0.2131	−2.950
Sweden	2.9687	0.7193	2.736
Switzerland	−16.2249	5.0798	−3.390
United Kingdom	−17.4481	2.0614	−8.949
United States	1.0000	NA	NA
Yugoslavia	1.7798	0.1524	5.115

Note: Number of observations = 297.
[a] Values are divided by U.S. estimate ($\delta_{us} = 1.0012$).
[b] Asymptotic t-values for testing δ_i is unity. The critical t-value based on equation (17) is 2.19.

may require a dramatic alteration of the H-O-V model. However, contrary to the suggested importance of these observations, re-estimation of the model for each hypothesis with the land variables excluded produced few changes in the estimated parameters (results not shown). Hypothesis H3 remained most favored, followed by hypothesis H7 and then HG. Under hypothesis H3, seventeen of the estimated technological differences δ_i exceeded unity and the number of countries with negative values of the technological difference parameter increased from eight to ten.[16] We thus remain confused about the exact source of the peculiar estimates.

The estimates reported in Table 7 are also cause for concern. The predicted values of the factor supplies can be found by inserting the observed values into these estimated equations. A negative value of γ_k indicates that the observed endowment and the "corrected" endowment are negatively correlated. This happens for four of the labor endowments, although three of these coefficients have large enough standard errors that the sign remains in doubt. This leaves production workers as the anomaly: the number of production workers embodied in trade is negatively related to the measured number of production workers.[17]

Overall, our results cast doubt on the hypothesis that the H-O-V equations are exact in favor of a model that allows neutral differences in factor input matrices and measurement errors in both trade and national resource supplies. This finding suggests that technological differences and measurement errors are also significant reasons for the relatively poor performance of the sharp hypotheses contained in the rank and sign propositions considered previously. However, our results do support the assumptions of proportional consumption[18] and complete coverage of countries. However, these conclusions are rendered suspect by the peculiar point estimates that are produced by the favored hypothesis.

TABLE 7—H-O-V REGRESSIONS AND FACTOR COEFFICIENTS UNDER HYPOTHESIS H3

Resource	Parameters α_k[a]	γ_k[b]
Capital	−990620794	13.431
	(−6.665)	(2.142)
Labor		
Agricultural	−7853	13.631
	(−1.376)	(2.721)
Clerical	−4628	−1.111
	(−1.426)	(−0.386)
Prof/Tech	−4376	−0.360
	(−1.866)	(−0.128)
Managerial	−1815	−0.528
	(−1.587)	(−0.370)
Production	−19608	−2.671
	(−1.997)	(−2.152)
Sales	−1214	0.216
	(−0.515)	(0.175)
Service	−1302	0.053
	(−0.498)	(0.052)
Land		
Arable	−2570651	1718.648
	(−62.891)	(52.545)
Forest	−2454843	833.206
	(−21.263)	(20.427)
Pasture	−202638	199.930
	(−2.275)	(9.163)

[a]Asymptotic t-values in parentheses. The critical t-value based on equation (17) is 2.19.
[b]Values of γ_k scaled by 10^3.

IV. Concluding Remarks

This paper has reported conceptually correct tests of the Heckscher-Ohlin proposition that trade in commodities can be explained

[16]The estimate for Belgium-Luxembourg switched from negative to positive, while the estimates for the Netherlands, Norway, and Spain switched from positive to negative.

[17]Parameter estimates for the unrestricted model HG were very similar to those reported for hypothesis H3. In particular, of the eight countries with negative values of δ_i in Table 6, only the value for France was positive. In addition, the signs and levels of significance of the parameters γ_k paralleled those shown in Table 7.

[18]This contrasts with Yutaka Horiba's (1979) test of the proportional consumption assumption using data on U.S. regional trade. Using a specification similar to ours, he rejected the assumption in terms of the value of β_k but not its sign.

in terms of an interaction between factor input requirements and factor endowments. An exact specification of this interaction in a multicountry, multicommodity, multifactor world was derived in the form of the Heckscher-Ohlin-Vanek (H-O-V) theorem, which equates the factors embodied in net trade to excess factor supplies. The H-O-V theorem was weakened to allow nonproportional consumption and technological differences and was supplemented with various assumptions about measurement errors. Using 1967 trade and input requirements, we tested the null hypothesis that the H-O-V equations are exact against several of these weaker alternatives. In addition, we examined sign and rank corollaries of the H-O-V theorem analogous to those implicitly studied by Leontief.

The Leontief-type sign and rank propositions, whether examined across countries or across factors, were generally not supported. The sign of net factor exports infrequently predicted the sign of excess factor supplies and therefore does not reliably reveal factor abundance. The ranking of factor contents infrequently conforms to the ranking of factor abundance ratios, as examined through either rank correlations or pairwise rankings.

The hypothesis that the H-O-V equations are exact was also not supported. The data suggest errors in measurement in both trade and national factor supplies, and favor the hypothesis of neutral technological differences across countries. However, the form of the technological differences favored by the data involves a number of implausible estimates, including some in which factors yield strictly negative outputs.[19] Thus, to a considerable extent, the conclusions that come from a study of the sign and rank propositions apply to the more promising regression study: The Heckscher-Ohlin model does poorly, but we do not have anything that does better. It is easy to find hypotheses that do as well or better in a statistical sense, but these alternatives yield economically unsatisfying parameter estimates.

These generally negative conclusions concerning the empirical validity of the H-O-V model appear to contrast sharply with Leamer's (1984) conclusion that "the main currents of international trade are well understood in terms of the abundance of a remarkably limited list of resources. In that sense the Heckscher-Ohlin theory comes out looking rather well." However, the present paper tests a different set of hypotheses. Leamer (1984) studies the weakened hypothesis that the structure of trade can be explained by the availability of resources. This paper examines the stricter H-O-V hypothesis that factor supplies, factor input requirements, and trade interact in a particular way. In addition, the present results suggest that there are important differences in selected input intensities between the United States and the other countries. Leamer's (1984) study may come to a more optimistic conclusion because he makes no commitment to the U.S. input intensities.[20]

DATA APPENDIX

Data on 1966 factor and 1967 trade endowments were collected for 27 countries. The twelve resources are capital, total labor, professional/technical workers, managerial workers, clerical workers, sales workers, service workers, agricultural workers, production workers, arable land, pastureland, and forestland. In accordance with this *Review's* policy of ensuring clear documentation of data, Table A1 lists the data on countries' population, GNP, and trade balance, as well as their trade and endowment of each factor. The following provides a concise discussion of data sources and methods, and includes citation to previously published work which contains further information on these data.

Factor endowment data were obtained from Bowen (1980, 1983). Net capital stocks for each country were

[19]Although the assumption of factor price equalization is not explicit in our analysis, the performance of hypothesis H3 together with the results shown in Table 6 could be taken as evidence against the assumption of factor price equalization. Factor price differences might help explain the variability in the estimates of δ_i since such differences would imply nonneutral differences in factor input matrices. We intend to examine the possibility of nonneutral technological differences in later research.

[20]See also Anderson's (1987) review of Leamer (1984).

TABLE A1—DATA BASE

(1)	(2)	(3)	Country		(4)	(5)	(6)	(7)
22.0	2.17431E10	4.68938E8	Argentina	E	2.40180E10	8496000	1342368	953251
				T	8.67853E8	77647	79417	1701
11.6	2.27360E10	−1.16039E8	Australia	E	3.50530E10	4727000	461828	702905
				T	−1.45799E9	−44830	78226	−14787
7.3	1.01688E10	−6.44060E8	Austria	E	1.56530E10	3363000	605340	393135
				T	−1.07484E9	−40112	−10955	−4982
9.8	1.89645E10	−3.45010E8	Bene-Lux	E	2.25630E10	3764000	236379	477652
				T	−9.38096E8	−7454	−26037	293
83.9	2.90170E10	2.45233E8	Brazil	E	3.04760E10	26463000	12696947	1267578
				T	−1.39899E9	−18704	6813	−4128
20.0	5.68412E10	4.24574E8	Canada	E	7.65370E10	7232000	690656	984998
				T	1.88821E9	−158932	47260	−28570
4.8	1.11664E10	−5.88141E8	Denmark	E	1.30180E10	2230000	304618	249760
				T	−1.32815E9	805	47564	−4358
4.6	8.64730E9	−2.20552E8	Finland	E	1.39290E10	2176000	574029	173862
				T	3.94420E8	−1796	−3009	−2530
49.2	1.08118E11	−9.53400E8	France	E	1.46052E11	21233000	3709405	2299534
				T	−7.06224E8	−34336	−38160	−736
59.7	1.22675E11	2.11160E9	Germany	E	1.81079E11	26576000	2854262	4217611
				T	5.72919E7	349286	−240808	75288
8.6	6.72160E9	−8.16871E8	Greece	E	7.22300E9	4314000	2065975	250212
				T	−1.35709E9	−52749	7616	−7373
3.6	1.97080E9	−4.43437E8	Hong Kong	E	2.08700E9	1525000	78842	104462
				T	−1.48229E9	−28429	−31726	−2834
2.9	2.93960E9	−3.84555E8	Ireland	E	3.37000E9	1109000	346895	90273
				T	−5.16958E8	−9731	18912	−3303

Note: See notes at end of table for column definitions.

Country		(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Argentina	E	588773	161424	2948112	927763	956650	30248000	60130000	145802000
	T	−998	−12	−6990	2886	1645	6561280	−3745350	3835996
Australia	E	443865	300164	2018429	369651	355470	39614000	35151000	447208000
	T	−15213	−6415	−82554	−1348	−2767	6762249	−4913655	3498840
Austria	E	260969	68941	1362688	261978	357487	1686000	3203000	2249000
	T	−4728	−1852	−14008	−1885	−1710	−1546894	−156822	87684
Bene-Lux	E	426461	145290	1778490	409900	275148	981000	687000	818000
	T	−2621	−436	22601	−805	−451	−3672691	−6653807	189337
Brazil	E	1098214	727732	5440793	1595719	2344622	31910000	522600000	141400000
	T	−4805	−1407	−14172	−646	−363	741997	15142	152291
Canada	E	868563	625568	2438630	496115	770208	43404000	322271000	20957000
	T	−21686	−10161	−134350	−5712	−5798	5389052	20253445	899259
Denmark	E	231474	35680	861226	224784	247530	2701000	472000	326000
	T	−5349	−2138	−34661	570	−824	737315	3254944	5328684
Finland	E	224998	32640	797286	165811	207590	2753000	21930000	110000
	T	−2960	−587	8710	−846	−579	−736283	6481551	320649
France	E	2344123	626373	7905046	1751722	1783572	20214000	12714000	13632000
	T	−1197	−1145	9304	−1872	−527	−4585442	−26133156	−437494
Germany	E	2479541	728182	11153947	2320085	2551296	8228000	7184000	5802000
	T	64071	28574	392143	13849	16244	−26018409	−25195922	−5730951
Greece	E	198444	29335	1134582	289901	309314	3851000	2608000	5239000
	T	−6174	−2893	−39697	−2046	−2187	1571541	−2343192	−526643
Hong Kong	E	77927	76097	755180	180102	247660	13000	10000	1
	T	−4095	−1153	14839	−1916	−1547	−2931651	−3459434	−1232976
Ireland	E	87278	14417	370739	109569	87611	1199000	208000	3554000
	T	−3175	−1431	−19559	−407	−769	96740	−622571	2308561

TABLE A1—CONTINUED

(1)	(2)	(3)	Country		(4)	(5)	(6)	(7)
52.0	7.96580E10	−5.39406E8	Italy	E	9.04360E10	19998000	4527547	1763824
				T	−6.98937E9	28922	−103119	11262
98.9	1.10388E11	2.53711E8	Japan	E	1.65976E11	49419000	11905037	6478831
				T	−8.78167E9	107388	−172142	27674
29.1	4.13000E9	−4.87021E8	Korea	E	3.02500E9	9440000	4936176	370048
				T	−1.49530E9	−49392	−9346	−5712
44.1	2.21625E10	−5.61792E8	Mexico	E	2.16390E10	12844000	5878699	910640
				T	−8.28121E8	−50163	25306	−9625
12.5	2.08090E10	−1.26535E9	Netherlands	E	2.99410E10	4699000	388607	657390
				T	−2.85259E9	−60964	29917	−8221
3.8	7.65510E9	−8.39611E8	Norway	E	1.28830E10	1464000	223699	125758
				T	−1.70045E9	−102657	−6254	−11350
32.7	6.18600E9	−1.04685E8	Philippines	E	6.59700E9	12470000	6660227	379088
				T	−1.04236E9	−35637	4721	−5728
9.3	4.26650E9	−4.03294E8	Portugal	E	3.75700E9	3381000	1166445	196436
				T	−8.61326E8	−23677	−11321	−2897
32.0	2.82285E10	−2.31879E9	Spain	E	3.47920E10	11849000	3673190	940811
				T	−4.87847E9	−148941	−17365	−17993
7.8	2.35715E10	−3.01150E8	Sweden	E	3.15550E10	3450000	368805	375705
				T	−1.03381E8	−19339	−16365	−3741
6.0	1.50576E10	−6.65988E8	Switzerland	E	2.33150E10	2843000	263546	457723
				T	−2.11706E9	−48845	−32919	−870
54.7	1.06534E11	−2.55290E9	United Kingdom	E	1.10717E11	25396000	891400	3512267
				T	−1.72415E10	−400012	−283790	−16862
196.5	7.62700E11	4.34870E9	United States	E	7.85933E11	76595000	3707198	2515623
				T	5.76114E9	764413	258625	87377
19.7	8.60700E9	−3.55350E8	Yugoslavia	E	1.40230E10	8837000	4539567	451571
				T	−8.59292E8	−17986	4413	−4003

Country		(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Italy	E	1181882	185981	8233177	2225777	897910	15258000	6099000	5147000
	T	4838	3877	112909	−1250	435	−62445355	−26821582	−7174026
Japan	E	2826767	1433151	16940833	5757313	3553226	5839000	25400000	157000
	T	16998	9108	220165	1543	4057	−19862383	−67987781	−2956563
Korea	E	244496	75520	1666160	927008	465392	2293000	6656000	18000
	T	−5232	−2186	−23143	−1903	−1873	−1111149	−2404532	−130644
Mexico	E	624218	245320	2766598	1075043	1344767	26000000	78000000	72100000
	T	−9184	−3865	−48936	−1760	−2103	3061733	3920090	297403
Netherlands	E	535686	123114	1823682	467550	429019	946000	292000	1299000
	T	−4618	−3443	−71238	−1260	−2084	417979	−3260982	3394099
Norway	E	158551	50508	653676	122390	129271	841000	7300000	158000
	T	−9529	−4377	−64732	−2976	−3447	−1297301	6966388	443706
Philippines	E	448920	342925	2035104	746953	872900	8330000	14100000	830000
	T	−4561	−2158	−24600	−1722	−1595	842059	−1283864	−216082
Portugal	E	109206	24005	1103220	225513	273185	4070000	3400000	530000
	T	−3159	−990	−3460	−1060	−791	−1274623	473446	−224250
Spain	E	596005	98347	4397164	893415	1071150	20156000	13600000	12000000
	T	−15604	−7010	−79483	−5828	−5650	−1221046	−3755719	−1066016
Sweden	E	566835	77625	1390350	317745	334305	3158000	22794000	525000
	T	−2183	−926	7287	−2105	−1317	−2210020	6774374	36931
Switzerland	E	307613	48900	1205432	209245	307897	401000	981000	1778000
	T	1446	−278	−14405	−948	−876	−3652367	−4066582	−692250
United Kingdom	E	2450714	789816	12027546	2460872	3088154	7480000	1829000	12107000
	T	−3730	−6651	−62500	−15201	−11232	−24179333	−49416509	−12948250
United States	E	9911393	7284184	28799720	5108886	9275654	177550000	306850000	258000000
	T	102628	30444	235781	24053	25598	35789032	−69100970	−1108796
Yugoslavia	E	609753	98974	2212785	269528	534638	8266000	8812000	6450000
	T	−4187	−1595	−10469	−1032	−1117	−108893	−78267	664739

Notes: E = 1966 endowment. T = 1967 net trade in factor. Units of T are those of the corresponding endowment. Columns are: 1) Population (mil.); 2) GNP (1966; $US); 3) Trade Balance (1966, $US); 4) Capital (1966, $US); Labor: 5) Total; 6) Agricultural; 7) Clerical; 8) Professional/Technical; 9) Managerial; 10) Production; 11) Sales; 12) Services; Land (hectares): 13) Arable; 14) Forest; 15) Pasture.

computed by summing annual real gross domestic investment flows starting in 1949 with annual depreciation assumed to be 13.33 percent. The underlying investment data were derived from the World Bank's Economic and Social Data Bank tape and appear in the World Bank publication *World Tables*. Detailed discussion of the methods used to construct net capital stocks appears in Bowen (1982).

Labor endowments were derived from issues of the International Labour Office (ILO) publication *Yearbook of Labor Statistics*. The labor categories are those defined at the one-digit level of the ILO's International Standard Classification of Occupations (ISCO). Total labor is defined as a country's total economically active population. For each country, the number of workers in each ISCO category was computed by multiplying the share of a country's total labor belonging to a category times its total labor. Since occupational data are not regularly collected, the share of each labor type in each country in 1966 was derived from a time-series regression of the available share data against time. Bowen (1982) provides discussion of this method and presents the years for which occupational data were available for each country.

Land endowments were taken from issues of the Food and Agricultural Organization (FAO) publication *Production Yearbook*. The definitions of arable land, pastureland, and forestland are those used by the FAO.

The total content (direct plus indirect) of each factor embodied in net trade was calculated by premultiplying each country's net trade vector by a matrix of total factor input requirements. Total factor input requirements were calculated from data on direct and indirect factor input requirements for each industry according to the 367-order U.S. input-output table for 1967. Data on each country's trade in 1967 were obtained from the U.N. Trade Data Tapes at the four- and five-digit level of the SITC and concorded to the input-output sectors to perform the required vector multiplications. The concordances are available from the authors upon request.

On the production side, capital (plant, equipment, and inventories) input requirements were constructed from data prepared by the Bureau of Labor Statistics Economic Growth Project, which provided industry capital stock figures measured in 1958 dollars. Industry occupation requirements, measured in number of persons, were based upon the 1971 Survey of Occupational Employment and the 1970 Census of Population. These data were reclassified, to the extent possible, to be consistent with the one-digit occupational categories defined by ISCO. (It was often not feasible to translate industry skill requirements into the ILO definitions; white-collar employment in certain nontraded sectors was a particular difficulty.) Sveikauskas (1983, Appendix) and especially Sveikauskas (1984) provide a complete description of the factor requirements data and a detailed table listing the input requirements data that can be made available.

Land inputs were constructed from information contained in the U.S. input-output table. Arable land is defined as proportional to total purchases from I/O sector 2; pastureland as proportional to total purchases from I/O sector 1 and forestland as proportional to total purchases from I/O sector 3 (which includes fisheries). This method of measuring land (natural resource) inputs corresponds to a rent definition of quantity and has been used by Baldwin (1971) and Harkness (1978), among others.

Land input coefficients are measured in dollars, whereas land endowments are measured in hectares. To adjust for this difference in units of measurement, the net trade in each land type was deflated using an imputed price. The prices were derived by dividing the total value of each type of land input absorbed in producing total U.S. output in 1967 by the corresponding U.S. endowment of each type of land in 1966. The prices, in 1967 dollars, are: arable land, \$142.767 per hectare; pastureland, \$108.942 per hectare; forestland, \$5.688 per hectare.

REFERENCES

Anderson, James E., "Review of *Sources of International Comparative Advantage: Theory and Evidence*, by Edward E. Leamer," *Journal of Economic Literature*, March 1987, *25*, 146–47.

Baldwin, Robert E., "Determinants of the Commodity Structure of U.S. Trade," *American Economic Review*, March 1971, *61*, 126–46.

Bowen, Harry P., *Resources, Technology and Dynamic Comparative Advantage: A Cross-Country Test of the Product Cycle Theory of International Trade*, unpublished doctoral dissertation, UCLA, 1980.

______, "Statistical Appendix to *Sources of International Comparative Advantage*," published, in part, as Appendix B of *Sources of International Comparative Advantage*, by Edward E. Leamer, Cambridge: MIT Press, 1982.

______, "Changes in the International Distribution of Resources and Their Impact on U.S. Comparative Advantage," *Review of Economics and Statistics*, August 1983, *65*, 402–14.

Branson, William and Monoyios, Nicholas, "Factor Inputs in U.S. Trade," *Journal of International Economics*, May 1977, *7*, 111–31.

Brecher, Richard and Choudhri, Ehsan, "The Leontief Paradox, Continued," *Journal of Political Economy*, August 1982, *90*, 820–23.

Chenery, Hollis and Syrquin, Moses, *Patterns of Development, 1950–1970*, New York and

London: Oxford University Press for World Bank, 1975.

Harkness, Jon, "Factor Abundance and Comparative Advantage," *American Economic Review*, December 1978, *68*, 784–800.

______, "The Factor-Proportions Model with Many Nations, Goods and Factors: Theory and Evidence," *Review of Economics and Statistics*, May 1983, *65*, 298–305.

Horiba, Yutaka, "Testing the Demand Side of Comparative Advantage Models," *American Economic Review*, September 1979, *69*, 650–61.

Kohler, Wilhelm K., "A Note on the Meaning of Leontief-Type Paradoxa," unpublished paper, 1987.

Leamer, Edward E., *Sources of International Comparative Advantage: Theory and Evidence*, Cambridge: MIT Press, 1984.

______, "The Leontief Paradox Reconsidered," *Journal of Political Economy*, June 1980, *88*, 495–503.

______, *Specification Searches*, New York: Wiley & Sons, 1978.

______, "The Commodity Composition of International Trade in Manufactures: An Empirical Analysis," *Oxford Economic Papers*, November 1974, *26*, 350–74.

______ **and Bowen, Harry P.,** "Cross-Section Tests of the Heckscher-Ohlin Theorem: Comment," *American Economic Review*, December 1981, *71*, 1040–43.

Leontief, Wassily, "Domestic Production and Foreign Trade: The American Capital Position Re-Examined," *Proceedings of the American Philosophical Society*, 1953, *97*, 332–49.

Maskus, Keith V., "A Test of the Heckscher-Ohlin-Vanek Theorem: The Leontief Commonplace," *Journal of International Economics*, November 1985, *9*, 201–12.

Schwarz, G., "Estimating the Dimension of a Model," *Annals of Statistics*, 1978, *6*, 461–64.

Stern, Robert M. and Maskus, Keith V., "Determinants of the Structure of U.S. Foreign Trade, 1958–76," *Journal of International Economics*, May 1981, *11*, 207–24.

Summers, Robert, Kravis, Irving and Heston, Alan, "International Comparisons of Real Product and Its Composition: 1950–1977," *The Review of Income and Wealth*, March 1980, *26*, 19–66.

Sveikauskas, Leo, "Science and Technology in United States Foreign Trade," *Economic Journal*, September 1983, *93*, 542–54.

______, "Science and Technology in Many Different Industries: Data for the Analysis of International Trade," *Review of Public Data Use*, June 1984, 133–56.

Vanek, Jaroslav, "The Factor Proportions Theory: The *N*-Factor Case," *Kyklos*, October 1968, *21*, 749–55.

International Labour Office, *Yearbook of Labor Statistics*, Geneva: International Labour Office, various years.

U.S. Department of Labor, Bureau of Labor Statistics (1974), *Survey of Occupational Employment*, Series of Reports 430, Occupational Employment in Manufacturing, 1971, Washington, D.C.

U.S. Bureau of the Census, U.S. Department of Commerce, *1970 Census of Population*, Occupation by Industry. PC(2)-7C, Washington: USGPO, 1972.

World Bank, *World Tables*, Baltimore: Johns Hopkins University Press for the World Bank, various years.

Vertical Product Differentiation and North-South Trade

By HARRY FLAM AND ELHANAN HELPMAN*

We develop a model of North-South trade in which the North exports high-quality and the South exports low-quality industrial products. Faster technical progress in the southern industrial sector leads the North to introduce new high-quality products and the South to abandon low-quality products. Production of northern low-quality products is shifted to the South. We also study the effects of technical progress in the North and population growth.

Economic progress is typically associated with the appearance of new products and the disappearance of old ones, with the former dominating the latter in available characteristics. There are countries that are the first to produce the new products and there are others that end up producing them after a time lag. Existing theories of international trade do not provide satisfactory explanations of these features, including Raymond Vernon's original product cycle hypothesis (1966) and its reformulation by Paul Krugman (1979) in a framework with horizontally differentiated products.[1]

We suggest an alternative model of North-South trade, which is based on vertical product differentiation; that is, differentiation according to quality. It predicts interesting patterns of trade dynamics as a result of population growth and technical progress. In particular, it predicts the appearance of new, high quality products, and the disappearance of old, low quality products (see also Nancy Stokey, 1986, on this point). It also predicts a quality-based product cycle; that is, when the North shifts production to higher quality varieties, it abandons the production of lower quality products whose production is taken up by the South. The structure of international trade is determined by cross-country differences in technology, income, and income distribution.

The North produces and exports high quality, high cost varieties, while the South exports low quality, low cost varieties. Given an overlap in income distribution, there exists intraindustry trade.[2] We analyze secular trends in intersectoral and intraindustry trade, in the available products, and the quality-based product cycle. Our analysis sheds light on the Burenstam Linder hypothesis (1961), because, like Rodney Falvey and Henryk Kierzkowski (1987), we assign a central role to income distributions.

We develop our model in Section I. In Section II we discuss the effects of changes in income distribution. This provides useful information about economic growth, because growth is typically associated with shifts in income distribution. However, since we do not have an explicit mechanism that links income distribution to growth, our

*Institute for International Economic Studies, University of Stockholm, S-106 91 Stockholm, Sweden, and Department of Economics, Tel-Aviv University, Tel-Aviv 69 978, Israel, respectively. We are grateful for financial assistance from the Bank of Sweden Tercentenary Foundation and the Foerder Institute for Economic Research. An earlier version of the paper circulated under the title "Trade Dynamics." We have benefited from comments from seminar participants at Haifa, Konstanz, Stockholm, Tel-Aviv, and Toronto.

[1] This applies to the recent extensions of Krugman's work by David Dollar (1986) and Richard Jensen and Marie Thursby (1986).

[2] This is also a feature of the model by Falvey and Kierzkowski (1987), which is a close kin of ours. Following Falvey and Kierzkowski, we define two-way trade in vertically differentiated products as intraindustry trade. Intraindustry trade is usually referred to in the literature as two-way trade in horizontally differentiated products (see, for example, Chapter 8 of Helpman and Krugman (1985)). However, in the type of model employed in this paper it is reasonable to associate different qualities of the same good with a single industry, and therefore also to identify two-way trade in different qualities as intraindustry trade.

analysis of this issue is not complete. Section III is devoted to the study of changes in the pattern of trade that result from population growth, and in particular from a differential rate of population growth, with the rate being higher in the less developed country. Technical progress is studied in Section IV. Here we emphasize a catch-up process in which the rate of technical change is faster in the less developed country. Section V is a summary.

I. The Model

It is assumed that two commodities exist: a homogeneous product and a vertically differentiated product. The homogeneous product can be consumed in every desirable quantity, whereas the consumption level of the differentiated product is fixed at unity. However, the consumer can choose the quality of the differentiated product from those available in the market. Consumer preferences are represented by a quasi-concave utility function $u(y, z)$, where y is the quantity of the homogeneous product and z is the quality of the differentiated product. Larger values of z represent higher quality. Therefore, $u(\cdot)$ is increasing in both arguments.

All individuals are identical except for income levels. An individual with income I choses a consumption level of the homogeneous product and a quality level of the differentiated product to solve the following problem:

$$\text{(1)} \qquad \max u(y,z) \text{ s.t. } y+p(z)\leq I \qquad y\geq 0 \quad z\in Z,$$

where $p(z)$ is the price of quality z, the price of the homogeneous product is one, and Z is the set of qualities available in the market. If the solution to this problem results in a utility level that is higher than the utility level that obtains from consuming only the homogeneous product, then the individual consumes both goods. Otherwise, the person consumes only the homogeneous product. We conduct the analysis under the assumption that every consumer finds it desirable to consume both products.

Two countries exist: a home country (North) and a foreign country (South). One unit of labor produces one unit of the homogeneous product in both countries. However, labor input per unit output of the differentiated product differs across countries. Let $a(z)$ and $a^*(z)$ be labor input per unit output of quality z in the North and South, respectively. These functions are convex and increasing in z. The North has comparative advantage in high quality products; that is, $a(z)/a^*(z)$ is declining in z. Now, assuming that the South produces the homogeneous product, its wage rate is equal to one (in terms of the homogeneous product) and the North's wage rate w is at least as large as one. The supply price of quality z is (see Sherwin Rosen, 1974):

$$\text{(2)} \qquad p(z)=\min[wa(z),a^*(z)].$$

This is also the price profile in a competitive equilibrium. Given the structure of comparative advantage, (2) implies that the South is the supplier of low quality products and the North is the supplier of high quality products. The break-even point in the chain of comparative advantage is a quality $\bar{z}$ that satisfies $wa(\bar{z})=a^*(\bar{z})$ (see Rudiger Dornbusch et al., 1977).

The consumer problem (1) can be represented graphically, as in Figure 1. The budget curve is $y=I-p(z)$. There is a set of usually shaped indifference curves representing $u(y,z)$. The consumer chooses a combination of (y,z) on the budget curve at the point of tangency with an indifference curve such as point A. An individual with a higher income faces a higher budget curve.

In what follows we use specific functional forms of the utility and unit labor input functions. The following equations prove to be convenient:

$$\text{(3)} \qquad u(y,z)=ye^{\alpha z}$$

$$\text{(4)} \qquad a(z)=e^{\gamma z}/A$$

$$\text{(5)} \qquad a^*(z)=e^{\gamma^* z}/A^*$$

with $\alpha>0$ and γ^*, $\gamma>0$. The North has

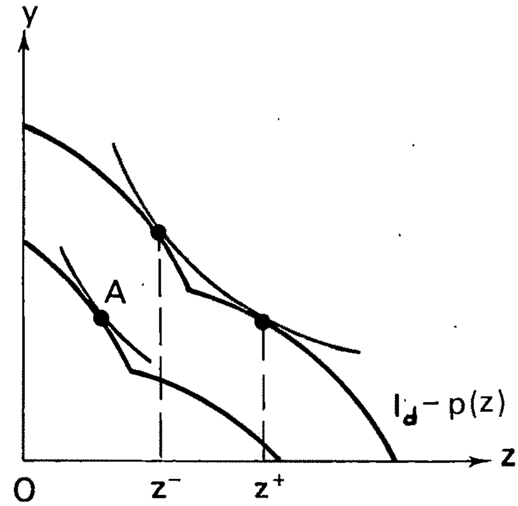


FIGURE 1

comparative advantage in high quality products if and only if $\gamma^* > \gamma$.

The utility function (3) has the property that the marginal rate of substitution between z and y depends only on y. Hence, the income expansion path for a given level of the marginal rate of substitution in consumption is horizontal in Figure 1. This implies that individuals with higher income consume more of the homogeneous product and a higher quality differentiated product (this is, of course, a feature of many other utility functions as well). Hence, if there exists an income level at which a southern-produced quality is demanded, and a higher income level at which a northern-produced quality is demanded, then there exists an intermediate income level—denoted by I_d —at which the consumer is indifferent between the consumption of a southern-produced quality z^- and a northern-produced quality z^+. The choice problem of a consumer with income I_d is also depicted in Figure 1. It is clear from this analysis that no demand exists for qualities in the range (z^-, z^+). This type of phenomenon appears also in previous studies (for example, Elhanan Helpman, 1985; Falvey and Kierzkowski, 1987).

Using the first-order conditions for problem (1), the functional forms (3)–(5), and the pricing equation (2), we obtain

$$(6)\qquad I = we^{\gamma z}\left(1+\frac{\gamma}{\alpha}\right)\Big/A \quad \text{for } I \geq I_d$$

$$(7)\qquad I = e^{\gamma^* z}\left(1+\frac{\gamma^*}{\alpha}\right)\Big/A^* \quad \text{for } I \leq I_d$$

$$(8)\qquad z^+ = \frac{1}{\gamma}\left[\log\frac{\alpha}{\alpha+\gamma} + \log I_d + \log A - \log w\right]$$

$$(9)\qquad z^- = \frac{1}{\gamma^*}\left[\log\frac{\alpha}{\alpha+\gamma^*} + \log I_d + \log A^*\right].$$

Since $p(z) = w\exp(\gamma z)/A$ for $z \geq z^+$ and $p(z) = \exp(\gamma^* z)/A^*$ for $z \leq z^-$, equation (6) implies that individuals with income above I_d—who buy northern-produced varieties —spend a share $\alpha/(\alpha+\gamma)$ of income on the differentiated product, while individuals with income below I_d—who buy southern-produced varieties—spend a share $\alpha/(\alpha+\gamma^*)$ of income on the differentiated product. This feature of our demand system makes it most convenient for the applications that follow.

Now, from the definition of I_d, it satisfies $u[I_d - p(z^+), z^+] = u[I_d - p(z^-), z^-]$, which with the help of (2)–(5) and (8)–(9) yields

$$(10)\qquad I_d^{\alpha(1/\gamma - 1/\gamma^*)} = Bw^{\alpha/\gamma}\left(A^{*\alpha/\gamma^*}/A^{\alpha/\gamma}\right),$$

where

$$B \equiv \frac{\gamma^* \alpha^{\alpha/\gamma^*}(\alpha+\gamma)^{(\alpha+\gamma)/\gamma}}{\gamma\alpha^{\alpha/\gamma}(\alpha+\gamma^*)^{(\alpha+\gamma)/\gamma^*}}.$$

Equation (10) describes the equilibrium relationship between I_d and the North's wage rate (the South's wage rate is equal to one). This relationship depends on the productivity parameters A and A^*, a feature that will be explored in our discussion of trade dynamics.

Every country is populated by a continuum of individuals, and a nondegenerate distribution of skills in the population exists. Differences in skills are reflected in differences in the endowment of effective labor supply. This is represented by means of in-

come classes. The set of income classes is chosen to be the unit interval [0,1]. The distribution of effective labor units across income classes is described by the density function $f(h)$ in the North and $f^*(h)$ in the South. That is, if L stands for the quantity of labor available to the North, then $f(h)Ldh$ of labor is supplied by northern individuals in income classes $[h, h+dh)$, $h \in [0,1)$, and similarly in the South. The distribution of the population over income classes is represented by the functions $n(h)$ and $n^*(h)$ and population sizes are N and N^*. Hence, the income level of a northern individual in income class h is $f(h)wL/n(h)N$, and similarly for the South. We choose higher values of h to represent higher income classes. It is therefore assumed that $f(h)/n(h)$ and $f^*(h)/n^*(h)$ are increasing in h.

In the remaining part of this section we describe an equilibrium in which the homogeneous product is produced only in the South, both countries produce some varieties of the differentiated product, and some of both countries' varieties are consumed in each one of them. We refer to this as the *central case*. Other patterns of specialization and consumption will be discussed in the sequel.

Since northern- and southern-produced varieties are consumed in both countries, there exists an income class h_d in the North and an income class h_d^* in the South, such that individuals who belong to them earn precisely I_d. Hence,

$$I_d = \frac{wLf(h_d)}{Nn(h_d)}, \tag{11}$$

$$I_d = \frac{L^*f^*(h_d^*)}{N^*n^*(h_d^*)}. \tag{12}$$

Northern individuals in income classes $(h_d, 1]$ and southern individuals in income classes $(h_d^*, 1]$ consume northern-produced differentiated products, and each one of them spends a proportion $\alpha/(\alpha+\gamma)$ of personal income on the differentiated product. Hence, total spending on northern varieties is the share $\alpha/(\alpha+\gamma)$ of the aggregate income of these two groups, which is $[1-F(h_d)]wL + [1-F^*(h_d^*)]L^*$, where $F(\cdot)$ is the cumulative distribution function associated with $f(\cdot)$ and similarly for $F^*(\cdot)$. Therefore, equilibrium in the northern labor market requires

$$\frac{\alpha}{\alpha+\gamma}\{[1-F(h_d)]wL + [1-F^*(h_d)]L^*\} = wL,$$

which reduces to

$$wL[\gamma + \alpha F(h_d)] = \alpha L^*[1-F^*(h_d^*)]. \tag{13}$$

The same condition can be derived from labor market clearing in the South or from the balancing of the trade account.

Equations (10)–(13) constitute a set of equilibrium conditions that determine I_d, w, h_d, and h_d^*. Having the values of those variables, one can use (8) and (9) to calculate z^- and z^+, and (6) and (7) to calculate the varieties consumed by individuals who have different income levels. Of particular interest is the range of varieties consumed by each country. For this purpose, we need to calculate the maximal and the minimal quality consumed in each country. Since the lowest quality is consumed by income class $h=0$ and the highest quality is consumed by income class $h=1$, (8) and (9) imply:

$$z_{\max} = \frac{1}{\gamma}\left[\log\frac{\alpha}{\alpha+\gamma} + \log[wLf(1)/N] + \log A - \log w\right] \tag{14a}$$

$$z_{\min} = \frac{1}{\gamma^*}\left[\log\frac{\alpha}{\alpha+\gamma^*} + \log[wLf(0)/N] + \log A^*\right] \tag{14b}$$

$$z_{\max}^* = \frac{1}{\gamma}\left[\log\frac{\alpha}{\alpha+\gamma} + \log[L^*f^*(1)/N^*] + \log A - \log w\right] \tag{15a}$$

$$z_{\min}^* = \frac{1}{\gamma^*}\left[\log\frac{\alpha}{\alpha+\gamma^*} + \log[L^*f^*(0)/N^*] + \log A^*\right]. \tag{15b}$$

The range of qualities purchased by northern consumers is $[z_{\min}, z^-] \cup [(z^+, z_{\max}]$ and the range of qualities purchased by southern consumers is $[z^*_{\min}, z^-] \cup [z^+, z^*_{\max}]$.

All these calculations are, of course, valid only if the patterns of specialization and consumption that we have chosen are equilibrium patterns. Namely, only if the underlying parameters and distribution functions imply that the values of (I_d, w, h_d, h_d^*), which solve (10)–(13), satisfy $w > 1$ and $0 < h_d,\ h_d^* < 1$. This is the central case; we will consider other patterns of consumption and specialization in due course.

In the central case the pattern of trade is as follows: The North exports high quality differentiated products and imports low quality differentiated products as well as the homogeneous product. Since trade is balanced, the volume of trade—defined as the sum of all exports—is twice the exports of a single country. In particular, it equals twice northern exports. Due to the fact that the South imports only differentiated products, that only individuals in income classes above h_d^* purchase imported varieties, and that each one of these individuals spends a share $\alpha/(\alpha+\gamma)$ of personal income on the differentiated product, the volume of trade can be represented by

$$(16)\quad VT = 2\frac{\alpha}{\alpha+\gamma}[1-F^*(h_d^*)]L^* = \frac{2}{\alpha+\gamma}[\gamma+\alpha F(h_d)]wL,$$

where the last equality derives from (13).

The volume of intraindustry trade is calculated in the usual way, as twice the sum across industries of the minimum across countries of exports of differentiated products. The fact that the North exports only differentiated products, the South exports homogeneous and differentiated products, and trade is balanced, implies that the volume of intraindustry trade equals twice southern exports of differentiated products. Varieties produced in the South are consumed by northern individuals whose income is not larger than I_d; that is, by individuals in income classes $h \le h_d$. Each one of these individuals spends a share $\alpha/(\alpha+\gamma^*)$ of personal income on the differentiated product. Hence, the volume of intraindustry trade is $2\alpha F(h_d)wL/(\alpha+\gamma^*)$, and the share of intraindustry trade is (using (16))

$$(17)\quad S_{i-i} = \frac{\alpha+\gamma}{\alpha+\gamma^*}\frac{wL}{L^*}\frac{F(h_d)}{1-F^*(h_d^*)} = \frac{\alpha+\gamma}{\alpha+\gamma^*}\frac{\alpha F(h_d)}{\gamma+\alpha F(h_d)} \le \frac{\alpha}{\alpha+\gamma^*}.$$

We see that the share of intraindustry trade depends on relative country size (as measured by relative GNP levels), on the income distribution in both countries, and the dividing income classes. The higher the relative income of the North and the larger the share of income of southern individuals that consume imported varieties, the larger is the share of intraindustry trade. Due to balanced trade, however, the share of intraindustry trade is larger, the larger the share of income of northern consumers that purchase imported differentiated products. The upper bound on the share of intraindustry trade is $\alpha/(\alpha+\gamma^*)$.

The comparative statics of the equilibrium conditions (10)–(13) can be derived by direct differentiation that yields the following linear system:

$$(18)\quad \begin{bmatrix} \alpha\left(\frac{1}{\gamma}-\frac{1}{\gamma^*}\right) & -\frac{\alpha}{\gamma} & 0 & 0 \\ 1 & -1 & -\varepsilon & 0 \\ 1 & 0 & 0 & -\varepsilon^* \\ 0 & 1 & \frac{\alpha f h_d}{\gamma+\alpha F} & \frac{f^* h_d^*}{1-F^*} \end{bmatrix} \times \begin{bmatrix} \hat{I}_d \\ \hat{w} \\ \hat{h}_d \\ \hat{h}_d^* \end{bmatrix} = \begin{bmatrix} d\pi_{10} \\ d\pi_{11} \\ d\pi_{12} \\ d\pi_{13} \end{bmatrix},$$

where ε is the elasticity of $f(\cdot)/n(\cdot)$ with respect to h evaluated at h_d, ε^* is the elasticity of $f^*(\cdot)/n^*(\cdot)$ with respect to h evaluated at h_d^*, a caret over a variable indicates a proportional rate of change, and $d\pi_i$, $i = 10, 11, 12, 13$, represents the exogeneous shift in equation (i). All of our comparative statics

exercises can be performed by means of (18) with an appropriate substitution of values for $d\pi_i$. For explicit calculations of many of the results given in the following sections, we refer to Harry Flam and Helpman (1986). The determinant of the matrix on the left-hand side of (18) is

$$(19)\quad \Delta_1 = \alpha\left(\frac{1}{\gamma} - \frac{1}{\gamma^*}\right)\varepsilon^*\varepsilon + \frac{\alpha}{\gamma^*}\frac{\alpha f h_d}{\gamma + \alpha F}\varepsilon^* + \frac{\alpha}{\gamma}\frac{f^* h_d^*}{1 - F^*}\varepsilon > 0.$$

II. Income Distribution Effects

In this section we consider the effect of income distribution on trade and production. In order to deal with this issue, suppose that the cumulative income distribution functions can be decomposed as follows:

$$(20a)\quad \tilde{F}(h) = F(h) + \mu G(h)$$

$$(20b)\quad \tilde{F}^*(h) = F^*(h) + \mu^* G^*(h),$$

where $F(\cdot)$ and $F^*(\cdot)$ are distribution functions, while the functions $G(\cdot)$ and $G^*(\cdot)$ equal zero at $h = 0$ and $h = 1$, they are nondecreasing on an interval of low values of h, and nonincreasing on its complement with higher values of h. Then, starting with $\mu = \mu^* = 0$, a small increase in μ represents a switch toward a more even income distribution in the North and a small increase in μ^* represents a switch toward a more even income distribution in the South.

We restrict our analysis to income redistribution schemes that shift income from income classes above the dividing ones, h_d and h_d^*, to income classes below the dividing ones. In this case, $G(\cdot)$ reaches a maximum at h_d and $G^*(\cdot)$ reaches a maximum at h_d^*. Assuming differentiability at these points, we have

$$\tilde{f}(h_d) = f(h_d) + \mu G'(h_d) = f(h_d) \quad \text{for all } \mu,$$

$$\tilde{f}^*(h_d^*) = f^*(h_d^*) + \mu^* G^{*\prime}(h_d^*) = f^*(h_d^*) \quad \text{for all } \mu^*.$$

Under these circumstances, starting with $\mu = \mu^* = 0$, the effects of changes in μ and μ^* can be described by means of (18) with

$$d\pi_{10} = d\pi_{11} = d\pi_{12} = 0,$$

$$d\pi_{13} = -\frac{G^*(h_d^*)}{1 - F^*(h_d^*)}d\mu^* - \frac{\alpha G(h_d)}{\gamma + \alpha F(h_d)}d\mu.$$

This implies a decline in the dividing income level, in the North's wage rate, in the dividing income class in the South, and in the dividing income class in the North. We have:

$$\hat{I}_d < 0,\ \hat{w} < 0,\ \hat{h}_d < 0,\ \hat{h}_d^* < 0.$$

These results stem from the fact that an income redistribution from individuals in income classes above the dividing ones toward individuals in income classes below the dividing ones shifts demand away from northern products, and toward the lower quality products that are produced in the South. Consequently, prices of varieties that are produced in the North decline, and so does its wage rate; the terms of trade move in favor of the South. Consumers face now a relatively lower price of qualities $z \geq z^+$, and those in income class h_d^* switch to a higher quality imported product. Hence, the dividing income class in the South declines. Northern consumers suffer a real income loss. Those in the dividing income class can mitigate some of this loss by switching to a now relatively cheaper home-produced, higher quality product, which they do, and the northern dividing income class declines.

In order to derive the effects of these changes on the share of intraindustry trade in the case in which income is redistributed only in the South, we use (20a) to calculate $d\tilde{F}(h_d)$. Then, since the dividing income class h_d declines and μ does not change, the share of northern income spent on imported differentiated products decreases. Consequently, the share of intraindustry trade declines (see (17)). In other words, the equalization of income distribution in the South moves the terms of trade in favor of the South and decreases the share of intraindustry trade.

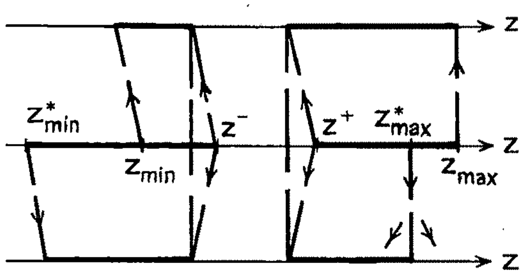


FIGURE 2

Figure 2 describes the shift in product spectrum that results from a more equal income distribution in the South. The middle line describes the initially assumed critical z values. The final values are described on the upper and lower lines. The North expands its product range by adding lower quality product lines. The South contracts its product range by abandoning both the highest and the lowest quality products.[3]

Finally, it should be observed that income redistribution within the open intervals $(0, h_d)$, $(h_d, 1)$, $(0, h_d^*)$ or $(h_d^*, 1)$ do not affect the variables that we have discussed. Thus, they do not change wages and dividing income classes. Therefore, they also do not change the share of intraindustry trade as well as production and consumption product ranges. They do, however, change quantities of products that are consumed by the affected income classes.

III. Population Growth

In this section we discuss trade dynamics that result from population growth. Consider the case in which the southern population increases at the same rate as the number of effective labor units, and at the same rate in every income class. Then we have $d\pi_{13} = \hat{N}^* = \hat{L}^* > 0$ and $d\pi_{10} = d\pi_{11} = d\pi_{12} = 0$. By substituting these values into (18), we find that

$$\hat{I}_d > 0, \quad \hat{w} > 0, \quad h_d > 0, \quad \text{and} \quad \hat{h}_d^* > 0.$$

Opposite effects result from a uniform population increase across income classes in the North. Moreover, when population increases at the same rate in both countries, there is no change in the structure of the equilibrium, except that output and trade volumes increase at the common rate of population growth. The share of intraindustry trade does not change.

When the southern population rises proportionately more than the northern population, demand for northern differentiated products rises by more than their supply. Consequently, the wage rate and output in the North increase, and the terms of trade move against the South.[4] These changes cause consumers in income class h_d^* to abandon consumption of the high quality variety that is produced in the North, and to switch to a locally produced lower quality product, whose relative price has fallen. They also cause consumers in income class h_d to abandon consumption of the domestically produced variety and to switch to a relatively cheaper, imported lower quality variety. The result is that the dividing income level I_d and the dividing income classes rise. It can be shown that the dividing income level rises proportionally more than the wage rate in the North. Hence, the lowest quality variety produced in the North, z^+, and the highest quality variety produced in the South, z^-, increase (see (8) and (9)). Since prices of varieties consumed by the lowest income classes in the South do not change, the variety consumed by the lowest income class z^*_{min} does not change as well (see (15)). Hence, the range of qualities produced in the South expands. On the other hand, the range of qualities produced in the North contracts, because the richest consumers in the North do not change the quality that they consume, z_{max}, whereas we have seen that the lowest

[3]The figure shows no change in z^*_{max}, which is not strictly correct; it may increase or decline, depending on how much income is taken away from $h^* = 1$ in the redistribution process. If, for example, no income is taken away from the richest individuals, then z^*_{max} increases due to the fact that these individuals' income increases in terms of the variety that they used to consume as well as in terms of higher quality products.

[4]This result is common to many models; it appears, for example, in Krugman (1979).

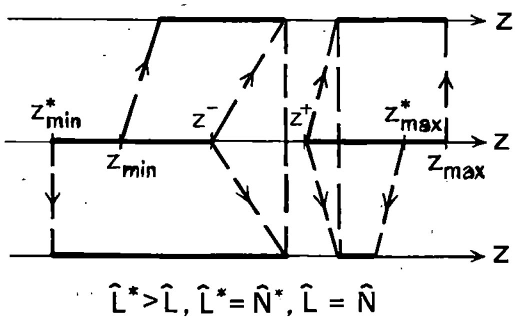


FIGURE 3

quality produced in the North is abandoned (see (9) and (14)). The resulting changes in the range of varieties produced and consumed in each country are represented graphically in Figure 3, where the initial ranges are described on the middle line. (The gap between z^+ and z^- is narrowed, as can be seen from equations (8) and (9).)

The cross-country distributional effects of this differential population increase are as follows. All northern consumers gain, southern consumers who are in income classes above h_d^* lose, and southern consumers who are in income classes below h_d^* are not affected. All northern consumers gain, because their income in terms of the homogeneous product increases. In addition, income in terms of the differentiated product increases for those in income classes below h_d and does not change for those in income classes above h_d. In the South income classes below h_d^* are not affected, because their income remains the same in terms of both the homogeneous and the differentiated product. On the other hand, income classes above h_d^* lose, because their income in terms of the homogeneous product does not change, and their income in terms of the differentiated product declines.

Finally, since h_d increases, northern consumers spend a larger share of income on imported varieties. This leads to an increase in the share of intraindustry trade (see (17)).

Our analysis implies the following trade dynamics. When the rates of population growth are the same in both countries, there is no change in the patterns of consumption, production, and trade. However, quantities consumed, produced, and traded are changing over time; all of them increase at the rate of population growth.

In the more interesting case in which the rate of population growth is higher in the South, we have seen that the northern wage rate, the dividing income classes in both countries, and the share of intraindustry trade, all increase over time. The dynamics of product ranges in consumption and production are described in Figure 3. Although product ranges change over time, there can be no qualitative change in the patterns of trade and production (see our working paper for a proof). Observe, however, that despite the fact that these patterns do not change, the range of qualities that are produced in every country and the range of qualities that are consumed in every country change over time. In particular, the North is continuously abandoning production of its lower quality products and the South is continuously adopting higher quality products. If the initial gap between z^+ and z^- is not too large, then the South eventually adopts varieties that were produced in the North. Hence, there emerges a product cycle for medium quality products, with a time lag between the cessation of production of a variety in the North and the adoption of the same variety in the South. This product cycle is different from Vernon's, but it is a product cycle nevertheless. Casual observation suggests that quality-based product cycles are important empirical phenomena.

IV. Technical Progress

Next, consider productivity changes. Productivity changes can come about in several ways. They can result from an overall labor productivity improvement that does not affect population size but changes effective labor supply, or they can be specific to the homogeneous product or to the differentiated product industry. In the latter case, they can be uniform across varieties or biased in favor of certain qualities. We will, of course, not analyze all of these possibilities, but rather choose some that are of particular interest.

The most interesting case seems to be technical progress in the industrial (differentiated product) sector; it permits an analysis of a widening or narrowing technology gap between the less developed and the advanced country. But this too can be done in two ways. One can assume that technical progress takes place through changes in the productivity parameters A and A^*, or through changes in the parameters γ and γ^*. In the former case the difference in comparative advantage is preserved, while in the latter case the degree of comparative advantage changes as well. Only the former case is discussed below, because the latter introduces complications without adding insights of comparable value.

In the case of uniform productivity increases in all varieties (increases in A and A^*), we have

$$d\pi_{10} = \alpha\left(\frac{\hat{A}^*}{\gamma^*} - \frac{\hat{A}}{\gamma}\right),$$

$$d\pi_{11} = d\pi_{12} = d\pi_{13} = 0.$$

By substitution of these values into (18), one obtains

$$\hat{w}/d\pi_{10} < 0, \quad \hat{h}_d/d\pi_{10} > 0, \quad \text{and}$$

$$\text{(21)} \quad \operatorname{sign}[\hat{f}_d/d\pi_{10}] = \operatorname{sign}[\hat{h}_d^*/d\pi_{10}] = \operatorname{sign}\left[\varepsilon - \frac{\alpha f h_d}{\gamma + \alpha F}\right].$$

Hence, an increase in southern productivity affects adversely the North's wage rate and induces the northern income class h_d to switch consumption from domestic to foreign varieties. These changes result from the fact that when southern productivity rises, southern prices of differentiated products decline, bringing about a demand shift toward southern varieties. Consequently, demand for northern varieties declines, and so does the demand for its labor. The result is a wage cut. However, northern wages do not decline in proportion to the South's productivity increase, so that after the adjustment relative prices of southern varieties remain lower. This induces some consumers in the North to switch to southern lower quality products.

Southern consumers experience an increase in income in terms of differentiated products because differentiated product prices decline in both countries. Whether this induces consumers just above the dividing income class h_d^* to switch consumption to southern varieties; that is, whether h_d^* increases or decreases, depends on income and price effects. If relative prices of southern-produced varieties fall to sufficiently low levels, that is, the wage rate w falls only slightly, consumers in income classes just above h_d^* switch to varieties produced in the South. If, on the other hand, relative prices of southern-differentiated products fall only slightly, then the income effect dominates, and consumers in income classes just below h_d^* switch to higher quality products. The precise condition for the sign on the change in h_d^* is given in equation (21).

Despite the ambiguity in the response of the dividing income class h_d^* to the productivity increase in the South, there is no ambiguity in the response of the share of intraindustry trade. Due to the fact that h_d increases, the share of intraindustry trade increases as well (see (17)). Thus, higher productivity (of the proportional type) in the southern industrial sector increases the share of intraindustry trade.

All the results that were reported so far are reversed when productivity increases proportionately in the northern industrial sector. Thus, for example, the share of intraindustry trade declines and northern wages increase. And if there is a simultaneous productivity increase in both countries such that $\hat{A}/\gamma = \hat{A}^*/\gamma^*$, then all these variables do not change. However, despite the mirror image response of these variables to southern- and northern-productivity changes, no such symmetry exists when it comes to quality ranges in production and consumption.

The shifts in quality ranges that result from a productivity increase in the South are described in Figure 4. The highest available quality does not change. However, the range of products produced in the North contracts

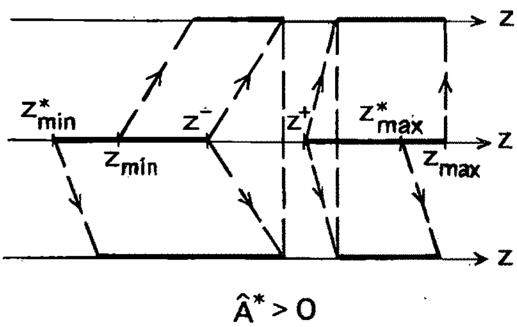


FIGURE 4

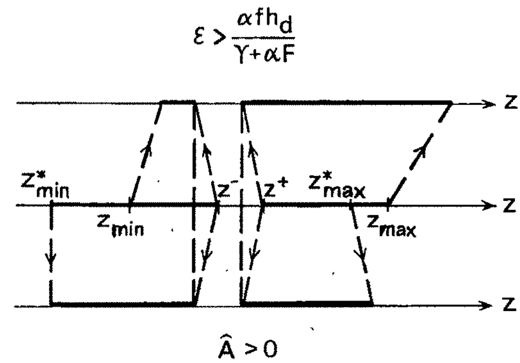


FIGURE 5

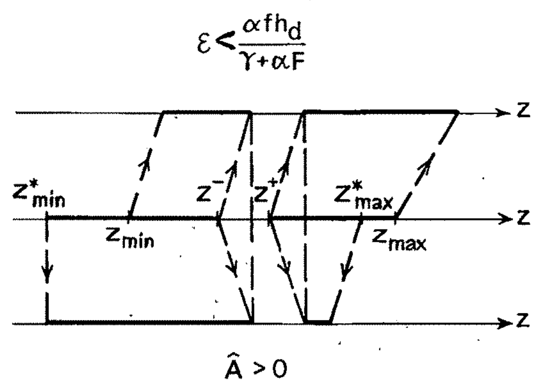


FIGURE 6

as a result of its abandoning the production of some low quality varieties, and the South also moves its production range up the quality spectrum; that is, to higher quality products. In both countries consumption shifts to higher quality varieties, except for the top ones in the North.

When productivity increases in the North, there are two possibilities, depending on whether the expression in (21) is positive or negative. If it is positive, the dividing income level and the dividing income classes decline, while the wage rate increases in the North. In this case, the increase in northern wages is proportionately smaller than the productivity increase, bringing about a fall in prices of differentiated products that are produced in the North. The resulting shifts in product ranges are described in Figure 5; the North produces a wider range of products, both of higher and lower quality, and the South produces a narrower range of products, abandoning the highest qualities. All northern consumers switch to higher quality products, while in the South only consumers in income classes above h_d^* switch to higher quality products. Other southern consumers do not change consumption patterns. Consequently, the spectrum of consumed varieties expands toward higher qualities.

When the sign in (21) is negative, one obtains the paradoxical result that northern wages rise proportionately more than the productivity increase. In this case, prices of differentiated products that are produced in the North *increase*, and so do the dividing income level and the southern dividing income class. The northern dividing income class declines as before; the income effect of a higher w dominates the relative price effect. The resulting changes in product ranges are described in Figure 6. The North shifts production to higher quality products and the South expands its product range by adding higher quality product lines. All northern consumers switch to higher quality products, while in the South consumers in income classes above h_d^* switch to lower quality products and consumers in income classes below h_d^* do not change their consumption pattern.

It is clear from this discussion that all northern consumers gain from this productivity increase, while (some) southern consumers gain only when the sign in (21) is positive; they lose when it is negative.

Our analysis implies the following dynamics. When the rate of technical change is

different across countries, but such that

$$\frac{\hat{A}}{\gamma}=\frac{\hat{A}^*}{\gamma^*}=\xi,$$

the rate of technical change is faster in the South, since we assume $\gamma<\gamma^*$. We saw in this case that the wage rate w, the dividing income level, and the dividing income classes do not change. Hence (8), (9), (14), and (15) imply that z^+, z^-, $z_{\max}$, $z_{\min}$, $z^*_{\max}$, and $z^*_{\min}$ all change at the rate ξ per unit time, so that all product ranges also shift upward at the rate ξ. Parallel to this uniform upgrading of qualities, every consumer upgrades at the rate ξ the quality that he consumes (see (6) and (7)). Thus, production and consumption move continuously to higher qualities at a uniform rate. In this steady state, new higher quality products are introduced in production by the North, and old low quality products are abandoned by the South. Also, the North abandons products that are adopted with a time lag by the South. A product cycle emerges again, but in this case it is a steady-state product cycle.

If technical progress takes place only in the South, then, as we have argued above, it leads to a decline in the North's wage rate and to an increase in its dividing income class. The South's dividing income class increases if and only if $\varepsilon>\alpha fh_d/(\gamma+\alpha F)$. The resulting shifts in product ranges are described in Figure 4. It is clear that as long as the central case equilibrium is preserved, production and consumption move to successively higher quality products in both countries except for the highest qualities. Here too a product cycle emerges; qualities that are abandoned by the North are adopted with a time lag by the South. However, once this process is set in motion, the increase in h_d and the decline in w will eventually lead to a change in the patterns of production and trade, because when the wage rate w becomes equal to one, the North becomes competitive in the production of the homogeneous product, and when h_d becomes equal to one, northern consumers cease to consume domestically produced differentiated products. On the other hand, the dividing income class h_d^* cannot reach its upper bound as long as h_d is positive and w is larger than one, as one can see from (13). Hence, two possibilities exist; either h_d becomes equal to one while w is larger than one, or w becomes equal to one while h_d is smaller than one. We only consider the latter possibility here; the former possibility is analyzed in our working paper.

In the case when w reaches one before h_d, the equilibrium conditions become (11) with $w=1$, (12), and

$$(10') \qquad I_d^{\alpha(1/\gamma-1/\gamma^*)}=B\frac{A^{*\alpha/\gamma^*}}{A^{\alpha/\gamma}},$$

$$(13') \quad L[\gamma+\alpha F(h_d)]-(\alpha+\gamma)Y = \alpha L^*[1-F^*(h_d^*)].$$

The four equilibrium equations determine changes in Y, I_d, h_d, and h_d^*, where Y is output of the homogeneous product in the North. Continuing technical progress in the South leads to increases in I_d, h_d, and h_d^*, and therefore also to higher output levels and lower imports of homogeneous goods in the North. The top qualities demanded, $z_{\max}$ and $z^*_{\max}$, remain constant (see (14) and (15)). At the same time, $z_{\min}$, $z^*_{\min}$, z^+, and z^- all increase (see also (8) and (9)). The range of differentiated products that are produced in the South moves up to higher qualities. The North abandons low quality products, and its range narrows down. Again we find a product cycle; that is, varieties that are abandoned by the North are adopted by the South after a time lag.

As the process continues, it is not h_d that reaches a value of one first, but rather h_d^*. This follows from (13'). Therefore, a new equilibrium is eventually reached in which the South has ceased to demand northern varieties, while the North is still demanding imported varieties. In the new equilibrium, the North exports homogeneous goods in return for low quality differentiated products. In fact, the pattern of trade is reversed at an earlier stage, when Y reaches the level

$$Y=\left[\frac{\gamma^*}{\alpha+\gamma^*}F(h_d)+\frac{\gamma}{\alpha+\gamma}[1-F(h_d)]\right].$$

If now technical progress continues in the South, then production of differentiated products will cease in the North; that is, h_d will reach one.

V. Summary

It has been demonstrated that our model of North-South trade, which incorporates vertical product differentiation, generates rich patterns of trade dynamics. A particularly appealing feature of these dynamics is that they explain the introduction of new high quality products and the disappearance of old low quality products, as well as the existence of a product cycle in which the less developed country begins to produce varieties that were produced in the advanced country. These are empirically relevant features.

Our model explains intertemporal shifts in intraindustry and intersectoral trade. Intraindustry trade arises because consumers who have different incomes demand different quality products, and because in a given country the range of produced qualities does not correspond precisely to the demanded range of qualities. The pattern of intraindustry trade reflects differences in technology and in income distribution; the South exports low quality, low cost varieties, while the North exports high quality, high cost varieties. This pattern is observed in many industries (for example, textiles, toys, radios).

Economic growth is typically associated with shifts in income distribution. We have not modeled this link. We did, however, analyze the pure effects of income redistributions. Equalization of the income distribution in the South, by shifting income from consumers who purchase high quality manufactures that are produced in the North to consumers who purchase low quality manufactures that are produced in the South, was shown to lower the wage rate in the North, to decrease the number of consumers that purchase manufactures from the South, and to decrease the share of intraindustry trade. The range of differentiated products that are produced in the South contracts, because it is led to abandon the production of its highest as well as lowest qualities. The North continues to produce its top quality varieties and adds product lines at its lowest quality end. Consequently, its product spectrum increases.

Faster population growth and faster technical progress in the South are probably the most interesting forces of economic growth that we have considered. We have shown that the former leads to a secular increase in the share of intraindustry trade and to a product cycle in middle-range quality products. However, it does not bring about changes in the available qualities at the lower and upper ends of the product spectrum. Hence, the pattern of trade does not change, although as time goes by fewer varieties are produced in the North and more varieties are produced in the South.

More rapid population growth in the South widens the income per capita gap between the North and the South. On the other hand, a higher rate of technical progress in the South narrows down this gap because it brings about a decline in the North's wage rate. The falling wage rate and the South's increasing relative efficiency in the production of industrial goods lead eventually to a switch in the patterns of production and trade. This is a necessary outcome of this process because: 1) the cost advantage of the South in production of homogeneous goods diminishes over time as a result of the fall in the North's wage rate, and 2) the South becomes successively more competitive in the production of differentiated products because the rate of technical progress proves to be more rapid than the fall of the North's wage rate. At the end of this process (if indeed it continues indefinitely), the North becomes specialized in the production of homogeneous goods and the South produces only differentiated products. However, before this final stage is reached, a product cycle exists, as the North is abandoning low quality varieties that are produced after a time lag in the South.

The above-described reversal in the patterns of production and trade can be reached along different trajectories, depending on the rate at which relative wages are falling. The major interest is in fact in the nature of these trajectories because they are relevant even

when the endpoint is never reached. Different trajectories are associated with different intermediate patterns of production and trade.

REFERENCES

Dollar, David, "Technological Innovation, Capital Mobility, and the Product Cycle in North-South Trade," *American Economic Review*, March 1986, *76*, 177–90.

Dornbusch, Rudiger, Fisher, Stanley and Samuelson, Paul A., "Comparative Advantage, Trade, and Payments in a Ricardian Model with a Continuum of Goods," *American Economic Review*, December 1977, *67*, 823–39.

Falvey, Rodney E. and Kierzkowski, Henryk, "Product Quality, Intra-industry Trade and (Im)perfect Competition," in H. Kierzkowski, ed., *Protection and Competition in International Trade*, Oxford: Basil Blackwell, 1987.

Flam, Harry and Helpman, Elhanan, "Trade Dynamics," Seminar Paper No. 354, Institute for International Economic Studies, University of Stockholm, 1986. Also Working Paper No. 25-86, Foerder Institute for Economic Research, Tel-Aviv University, 1986.

Helpman, Elhanan, "International Trade in Differentiated Middle Products," in Karl G. Jungenfelt and Douglas Hague, eds., *Structural Adjustment in Developed Open Economies*, London: Macmillan, 1985.

Helpman, Elhanan and Krugman, Paul, *Market Structure and Foreign Trade*, Cambridge: MIT Press, 1985.

Jensen, Richard and Thursby, Marie, "A Strategic Approach to the Product Life Cycle," *Journal of International Economics*, November 1986, *21*, 269–85.

Krugman, Paul, "A Model of Innovation, Technology Transfer, and the World Distribution of Income," *Journal of Political Economy*, April 1979, *87*, 253–66.

Linder, Staffan Burenstam, *An Essay on Trade and Transformation*, Uppsala: Almqvist & Wiksell, 1961.

Rosen, Sherwin, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, January-February 1974, *82*, 34–55.

Stokey, Nancy L., "Learning-by-Doing and the Introduction of New Goods," Discussion Paper No. 699, Center for Mathematical Studies in Economics and Management Science, Northwestern University, September 1986.

Vernon, Raymond, "International Investment and International Trade in the Product Cycle," *Quarterly Journal of Economics*, May 1966, *80*, 190–207.

Discretionary Trade Policy and Excessive Protection

By ROBERT W. STAIGER AND GUIDO TABELLINI*

This paper proposes a positive theory of tariff formation, based on the idea that the optimal trade policy may be time inconsistent. A benevolent government with redistributive goals may have an incentive to provide unexpected protection, since the redistributive effects of trade policy are larger if the policy is unanticipated. The suboptimal but time-consistent policy involves an excessive amount of protection. Furthermore, in a time-consistent equilibrium tariffs may dominate production subsidies. Thus, the requirement of time consistency can lead to a reversal of the traditional normative ordering of tariffs and subsidies as instruments of trade policy.

Two central conclusions of the pure theory of international trade are that a policy of free trade is optimal for a small country and that, if a protective policy is nonetheless adopted, a production or consumption subsidy would be preferable to a tariff or other trade-distorting policy. Equally central to the empirical record is the observation that active protectionist programs are widely pursued by countries with little or no apparent world market power, and that tariffs or other trade-distorting policies form the heart of such programs. Given the normative implications of the pure theory, an intriguing question raised by these empirical observations is why governments choose to do what they do.

In this regard, one line of research has brought into question the empirical relevance of the assumptions underlying the normative case for free trade, attacking both the small country and the perfect market assumptions as empirically unreasonable, especially for many developing countries. However, as Anne Krueger (1984) concludes in her review of this literature,

> By and large, theory and empirical evidence have combined to reassert the proposition that trade intervention is seldom optimal, even in the presence of market imperfections. It might even worsen the situation as contrasted with *laissez-faire*, especially if the intervention is too highly protectionist. [Krueger, p. 566]

A second line of research has focused on the political economy of protection, describing the pattern of protection either as a result of the government's politically motivated concern over the distribution of income across voters and powerful special interests, or as a means of achieving some exogenously postulated social objective (see Robert Baldwin, 1984a, and Wolfgang Mayer, 1984). In either case, a pattern of protection emerges which is at variance with the free trade implications of pure trade theory. Consequently, the political economy approach gives rise to a positive theory of protection that is at least potentially consistent with empirical observation. Moreover, once levels of protection are modeled as the outcome of a political process, tariffs can become socially preferable to production subsidies, provided there are sufficient differences in lobbying effectiveness under the two regimes (see Dani Rodrik, 1986).

*Stanford University, Stanford, CA 94305, and University of California, Los Angeles, CA, 90024, respectively. We have benefited substantially from the comments and suggestions of Alberto Alesina, Robert Baldwin, Alan Deardorff, Ronald McKinnon, Giovanna Mossetti, J. David Richardson, Michael Riordan, Dani Rodrik, Kenneth Rogoff, and Robert Stern, and members of workshops at the NBER Summer Institute, Stanford University, UCLA, The University of Michigan, and The University of Wisconsin. Any remaining errors are our own. An earlier version of this paper, entitled "The Time Inconsistency of Free Trade," was circulated as Memorandum No. 268 of the Center for Research in Economic Growth, Stanford University.

This paper suggests an additional determinant of patterns of protection, based on the idea that the optimal trade policy is time inconsistent, and hence may lack credibility with the owners of domestic factors of production. In the model we explore, it is the government's inability to precommit to an optimal policy of free trade that forces it to adopt a policy of protection, and which could lead it to prefer the use of tariffs over production subsidies.[1]

We consider a model of a small open economy in which tariffs are used by the government to redistribute income from individuals with a low marginal utility of income to those with a high marginal utility of income, subsequent to the realization of an adverse terms of trade shock. Jonathan Eaton and Gene Grossman (1985) explore the use of tariffs in response to terms-of-trade shocks in a model in which labor is perfectly mobile *ex ante* and *ex post*, but where capital is only mobile *ex ante*, that is, before the shock is observed. *Ex post* capital is completely immobile. In such a model they show that a protective policy can be optimal *ex ante*, in the sense that it can achieve some beneficial risk sharing between risk averse individuals. They also note that this policy is time inconsistent in the following respect. The expected tariff policy influences the sectoral allocation of capital *before* the shock is observed. Hence, the government may have an incentive to announce a policy different from the one that it would implement *ex post*, in an effort to affect the *pre-shock* allocation of resources. Eaton and Grossman provide some numerical simulations, indicating that the difference between the optimal and time-consistent policy is negligible.

In this paper we consider the behavior of the government and of the private sector *ex post*, after the shock is observed. Hence, the time inconsistency analyzed by Eaton and Grossman does not arise. We show however that if the post-shock sectoral reallocation of labor is costly, the implementation of a protective policy of risk sharing can suffer from a second (and potentially more important) time inconsistency. In the model considered below, a benevolent government has an incentive to surprise the private sector after the shock is observed by providing more protection than expected. The incentive to surprise is due to the fact that, if the protective policy is anticipated, it tends to reduce the amount by which labor reallocates away from the injured sector; as a result, production is less efficient and the redistributive impact of the tariff is diminished. Surprise protection is therefore a more effective and less costly means of redistributing income. However, the government will not be able to systematically surprise the private sector with more protection than expected. If commitments to an optimal but time-inconsistent policy are not credible, a time-consistent equilibrium will obtain in which the post-shock relocation of labor across sectors incorporates the expectation of trade restrictions, irrespective of the government announcements; and the government fulfills those expectations with a socially excessive level of protection.

In order to focus exclusively on time inconsistency as a cause of trade restrictions, we consider a situation in which free trade is the optimal (but time inconsistent) policy, and the government's inability to commit leads to a time-consistent policy of protection. With respect to a positive theory of tariff formation, the notion of time inconsistency suggests that governments with some degree of discretion in trade policy may find commitments difficult to make, and may often be forced to choose inferior, overprotective but time-consistent policies. This comes about not as a result of lobbying pressures or other political concerns asso-

[1] The main reference for the issue of time consistency of optimal economic policy is the seminal paper by Finn Kydland and Edward Prescott (1977). Kevin Roberts (1984) addresses some of the methodological points that arise in this paper. The notion of time inconsistency has also been applied to trade policy in Maskin and Newbery (1986). They have shown that, for a large importing country, the optimal tariff on an exhaustible resource is generally time inconsistent; the time-consistent tariff in their model can either exceed or fall short of the optimal. See also Arye Hillman, Eliakim Katz, and Jacob Rosenberg (1987).

ciated with the political economy literature. Rather, it is a consequence of the government's inability to precommit to trade policies that, *ex post*, it would not find optimal to pursue.

Moreover, the policy requirement of time consistency can lead to a reversal of the traditional normative ordering of tariffs and subsidies as instruments of trade policy: under certain conditions on the parameter values of the model, we show that a time-consistent tariff policy is preferred to a time-consistent production subsidy. This result can contribute to an explanation of the empirical puzzle that was noted above: that is, why protection might take the form of trade distortions rather than of production or consumption subsidies.

Finally, these theoretical results contain a clear normative implication for improving on the time-consistent but suboptimal equilibrium: the government should be enabled to undertake binding commitments concerning its future behavior. From an operational point of view, this is suggestive of the important role that could be performed by an international organization like the GATT: namely, to enforce the domestic commitments to a policy of free trade. The GATT was originally conceived to facilitate international cooperation among individual countries; the results of the paper suggest that this institution can—and presumably to some extent already does—perform an equally crucial role in enforcing the cooperative outcome in a setting in which the strategic interaction is between each country and its own domestic residents.

The remainder of the paper proceeds as follows. Section I presents the model within which our analysis will be carried out. Section II considers the role of expected protection in determining the relocation of labor across sectors. The optimal tariff policy under the assumption that the government can undertake a binding commitment is derived in Section III. The time-consistent policy is derived in Section IV, and the results are compared with those of the previous section. Section V explores conditions under which a time-consistent tariff policy would be preferred to a time-consistent production subsidy. A concluding discussion appears in Section VI.

I. The Model

We consider a small open economy with two traded goods and two inputs to production, labor and capital. Capital is immobile across sectors, and technology is homogeneous of degree one in both inputs. To simplify notation, we assume that both sectors share the same production function, and that each sector is endowed with one unit of capital. The production technologies are given by

$$(1) \qquad i = f(N^i), \qquad i = x, y,$$
$$f'(\cdot) > 0, \quad f''(\cdot) < 0,$$

where $N^i \equiv$ labor employed in sector i; $x \equiv$ exported good; and $y \equiv$ imported good.

In each sector, firms combine labor with their fixed stock of capital, up to the point where the value of labor's marginal product is equated to the nominal wage measured in terms of any numeraire, W^i:

$$(2) \qquad p^x f'(N^x) = W^x,$$
$$p^y(1+t) f'(N^y) = W^y,$$

where p^x is the world price of the exported good, t is the *ad valorem* tariff on imports, and p^y is the world price of imports. Wages are assumed to be perfectly flexible so that (2) yields the nominal wage that clears the labor market in each sector.

Throughout the paper, we will consider the reaction of private agents and of the government to a terms-of-trade shock that lowers the world price of good y by an amount ε. The case of a negative shock to the price of exports is symmetric, except that the government would choose an export subsidy rather than a tariff. Prior to the realization of ε: (i) labor is assumed to be allocated equally between the two sectors, with the initial sectoral employment normalized to unity; (ii) goods units are chosen so that the world prices of x and of y are equal, and their common prices normalized to unity; and (iii) free trade is assumed to prevail in the domestic country. Given this,

equation (2) implies that, prior to the terms-of-trade shock, nominal wages are equal in the two sectors: $W^x = W^y$.

The aggregate supply of labor in the economy as a whole is assumed fixed. Labor is mobile between sectors, and reallocates in response to the terms-of-trade shock on the basis of the expected intersectoral wage differential, W^{ye}/W^{xe}. A central feature of the model is that, in the absence of binding commitments, the government cannot irrevocably set tariff policy before the labor reallocation decision is made. In other words, the timing of the decisions after the shock is observed is that, either: (a) first workers reallocate and then a tariff is imposed; or (b) the labor reallocation and the tariff decisions are made simultaneously. Since the private sector is atomistic and takes the government tariff as given, there is no relevant distinction between (a) and (b): both are valid interpretations of the equilibrium presented below. Under either of these timing assumptions, and using (2), the expected wage differential subsequent to the shock is given by

$$(3)\quad \frac{W^{ye}}{W^{xe}} = \frac{(1-\varepsilon)(1+t^e)f'(N^y(t^e))}{f'(N^x(t^e))}.$$

Equation (3) makes clear that the expected wage differential depends on the expected tariff, t^e, since the actual tariff is observed only once the reallocation is completed. The relationship between W^{ye}/W^{xe} and t^e will be analyzed more thoroughly in the next section.

A crucial assumption of the model is that labor's intersectoral mobility comes only at a cost. Specifically, we assume that, whenever one unit of labor moves from one sector to the other, its marginal product falls by the fraction $(1-\lambda)$, $1>\lambda>0$. This hypothesis can be motivated, for instance, by the notion that each worker has to acquire some sector-specific human capital, through experience or through training, before it can become as productive as workers already employed in that sector. Thus one can interpret $(1-\lambda)$ as the fraction of labor time that a worker relocating from one sector to the other has to spend in acquiring new productive skills.[2] Since we consider a one-period model, the issue of how the marginal product of the newly hired workers evolves over time does not arise. Section VI briefly discusses how to extend this approach to an explicit intertemporal framework.

From this assumption it follows that in equilibrium the expected wage differential between the two sectors of the economy is constrained by the following inequalities,

$$(4)\qquad \frac{1}{\lambda} \geq \frac{W^{ye}}{W^{xe}} \geq \lambda.$$

If either inequality is violated, workers would find it optimal to move from one sector to the other, until condition (4) is satisfied. We abstract throughout the paper from corner solutions in which the entire labor force locates in a single sector.

The subsequent analysis will be simplified by expressing labor in efficiency units. Thus, letting α denote the fraction of labor that remains in sector y after all reallocation has taken place, the effective quantity of labor employed in each sector when the adjustment to the shock is completed is given by

$$(5)\qquad N^y = \alpha, \quad N^x = (1-\alpha)\lambda + 1.$$

The first equation in (5) follows from having normalized to unity the quantity of labor initially in each sector. The second equation incorporates the notion that N^x is expressed

[2]See Gary Becker, 1962, and Baldwin, 1984b, for a more detailed motivation of this assumption. Baldwin, 1982, contains a discussion of the incomplete nature of adjustment assistance programs in eliminating the private costs of intersectoral labor movements. Michael Mussa, 1982, and Joshua Aizenman and Jacob Frenkel, 1986, model imperfectly mobile labor in a more general way, allowing for a continuum of λ's that reflect individual-specific moving costs. Modeling imperfectly mobile labor in this way would alter the characterization of the optimal policy in our model, but would leave unaffected our basic result that the opportunity for discretion in trade policy leads to excessive protection. Finally, the formal analysis would be different, but the qualitative results would remain unchanged, if the mobility costs were modeled as subjective disutility costs rather than output losses, or as payments made to a third sector responsible for moving workers between sectors.

in efficiency units, and that labor which has relocated to sector x from y has a marginal product equal only to a fraction λ of the marginal product of labor already employed in x. The fraction α of labor that remains in sector y is determined endogenously in response to the expected wage differential. The determination of α is the focus of the next section.

Define I as national disposable income valued at domestic prices. Imposing the condition of balanced trade at world prices, abstracting from domestic taxes, and assuming that the tariff is nonprohibitive, it follows that, subsequent to the shock

$$(6) \quad I = f(N^x) + (1-\varepsilon)(1+t)f(N^y) + T,$$

where the tariff revenue T is defined by

$$(7) \quad T = t(C^y - f(N^y))(1-\varepsilon),$$

C^y being aggregate demand for the imported good y. Substituting (7) into (6), national income valued at domestic prices is given by

$$(8) \quad I = f(N^x) + (1-\varepsilon)f(N^y) + t(1-\varepsilon)C^y.$$

In order to focus on the redistributive consequences of tariffs for the labor allocation decision, we assume that the distribution of income is determined solely on the basis of the wage differential between the two sectors. Thus, income from capital and tariff revenues is distributed to each worker in proportion to the share of his labor income in the economywide wage bill.[3] Define the income share variable, ϕ, as

$$(9) \quad \phi^i \equiv \frac{I^i}{I} = \frac{W^i}{W^x + \alpha W^y + (1-\alpha)W^{yx}}, \quad i = x, y, yx,$$

where I^i is total disposable income of a worker of the ith type (valued at domestic prices), and the superscripts x, y, and yx denote workers originally in sector x who remain there, workers originally in sector y who remain there, and workers originally in sector y who move to sector x, respectively.

Each worker consumes a bundle of x and y, chosen so as to maximize an identical homothetic utility function, subject to a standard budget constraint. The indirect utility function of a representative consumer of the ith type is assumed to exhibit diminishing marginal utility of income, and can be written in terms of the previous notation as

$$(10) \quad V^i = V(p^x, p^y, I^i), \quad i = x, y, yx.$$

Letting V_p^i and V_I^i denote the partial derivatives of (10) with respect to p^y and I^i, respectively, the consumption of y on the part of consumers of the ith type can be expressed, using Roy's identity, as

$$(11) \quad C_i^y = -\frac{V_p^i}{V_I^i} = \phi^i C^y, \quad i = x, y, yx.$$

The second equality follows from the assumption that the common utility function is homothetic.

Finally, the government chooses a level of protection in response to the shock ε, so as to maximize a welfare function defined over the indirect utility functions of the three types of workers,

$$(12) \quad J = \alpha V^y + (1-\alpha)V^{yx} + V^x.$$

Thus, the weights in the social welfare function given to the three types of workers are chosen to be proportional to the size of the group to which each type belongs. A property of the model will be that, in equi-

[3] Hence, we are assuming that lump-sum instruments are available to distribute tariff revenue to consumers, but that these instruments are insufficiently flexible to achieve the optimal distribution of income across workers. In the extreme, as assumed here, the distribution of tariff revenues leaves the income distribution completely unaffected. If tariff revenues could be redistributed optimally, none of the qualitative results of the paper would change, except that the optimal tariff policy would in general involve some positive level of protection.

librium, $V^y = V^{yx}$ (that is, workers must be indifferent between staying in y or moving to x). Hence, the coefficient α in (12) affects the equilibrium value of J only indirectly, through its effect on the equilibrium trade policy, and not directly as a weighting factor. We mention the implications of more or less egalitarian weighting factors for the resulting trade policy in a later section.

We assume the absence of any market mechanisms for reallocating the risk associated with the terms-of-trade shock. If such private insurance markets existed and worked perfectly, there would be no role for government intervention in the form of protection in our model. Support for such an assumption is contained in Eaton and Grossman (1985).

II. The Reallocation of Labor Across Sectors

As discussed in the previous section, workers will respond to a fall in the world price of imports by reallocating across sectors in order to take advantage of any expected wage differential in excess of λ. We assume that the world price of y falls by an amount $\varepsilon > 1-\lambda$. This inequality assures that, if a zero tariff were expected by workers in sector y (that is, if $t^e = 0$), some sectoral movement of labor from y to x would take place.[4]

The movement of workers across sectors in response to ε will assure that, in equilibrium, the expected wage differential subsequent to the shock will satisfy the inequalities in (4). Making use of equations (3) and (5), the expected wage differential can be rewritten as

$$\frac{W^{ye}}{W^{xe}} = \frac{(1+t^e)(1-\varepsilon)f'(\alpha)}{f'([(1-\alpha)\lambda+1])}. \tag{13}$$

The fraction α of workers that remain in sector y after the shock is a function of t^e, defined implicitly by (13) and (4). The properties of α as a function of t^e are summarized in the following:

LEMMA: (*i*) *For* $1/(1-\varepsilon) \geq 1+t^e \geq \lambda/(1-\varepsilon)$, α *equals unity, and* W^{ye}/W^{xe} *is a continuous, differentiable, and strictly increasing function of* t^e, *with* $1 \geq W^{ye}/W^{xe} \geq \lambda$. (*ii*) *For* $\lambda/(1-\varepsilon) > 1+t^e \geq 1$, W^{ye}/W^{xe} *equals* λ, *and* α *is a continuous, differentiable, and strictly increasing function of* t^e, *with* $1 > \alpha > 0$.

PROOF:

For $\alpha = 1$, the expected wage differential W^{ye}/W^{xe} is determined by (13) as $W^{ye}/W^{xe} = (1+t^e)(1-\varepsilon)$. Therefore, if $1/(1-\varepsilon) \geq 1+t^e \geq \lambda/(1-\varepsilon)$ and $\alpha = 1$, equilibrium condition (4) is satisfied; given the initial allocation of labor, $W^{ye} \geq \lambda W^{xe}$. Consequently, when the condition in part (*i*) of the lemma is met, no worker finds it worthwhile to relocate from sector y to sector x. Finally, with $\alpha = 1$, the rest of part (*i*) follows immediately from (13). Alternatively, if $\lambda/(1-\varepsilon) > 1+t^e \geq 1$, then the second inequality in (4) will be violated if $\alpha = 1$; under the initial allocation of labor, $W^{ye} < \lambda W^{xe}$. Consequently, when the condition in part (*ii*) of the lemma holds, the second inequality in (4) will in equilibrium hold with equality for some $1 > \alpha > 0$, and labor will move from sector y to sector x until $W^{ye} = \lambda W^{xe}$. The rest of part (*ii*) then follows immediately by applying the implicit function theorem.

The lemma is illustrated in Figure 1. The horizontal axis measures the fraction α of workers remaining in sector y. The vertical axis measures the wage rate. Given the concavity of the production function, W^y is decreasing in α, and is represented for $t = 0$ by the downward sloping solid curve. Conversely, the wage that can be earned by moving to sector x, λW^x, is given by the upward sloping curve. If protection is neither anticipated nor forthcoming, the equilibrium allocation is given by $\alpha(t^e = 0)$, and corresponds to the point where the cost of reallocating is just equal to the wage differential, $W^y = \lambda W^x$. The imposition of a tariff shifts the W^y curve to the right, say to the dotted

[4] This can be seen by noting that, in the absence of any intersectoral labor movement, $N^x = N^y = 1$, and (3) implies $W^{ye}/W^{xe} = (1-\varepsilon) < \lambda$, which violates the labor market equilibrium condition (4).

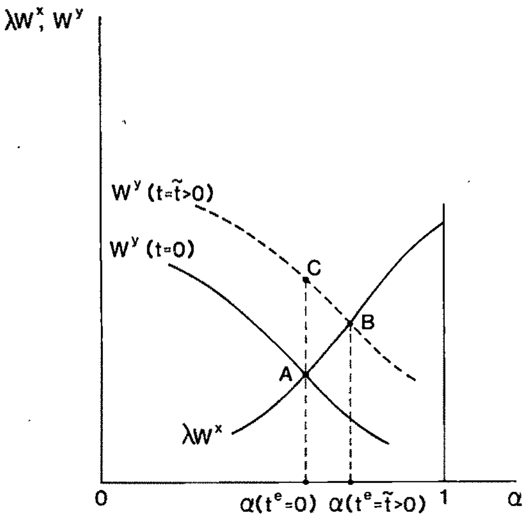


FIGURE 1

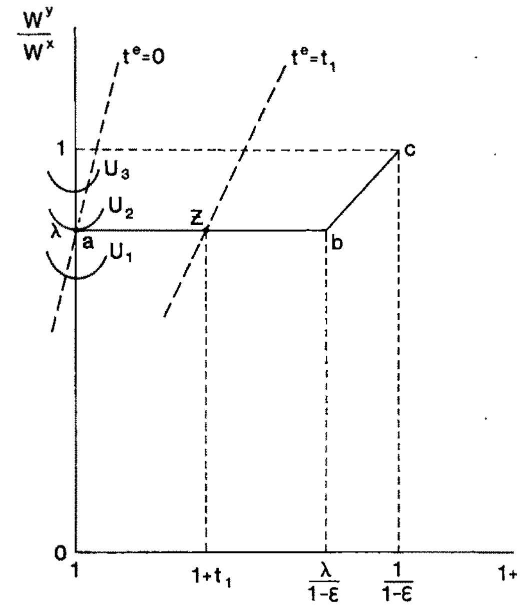


FIGURE 2

line of Figure 1. With the tariff fully anticipated, the equilibrium labor allocation is now given by point B in the diagram, where fewer workers have reallocated from sector y to x, that is, $\alpha(t^e=\tilde{t}>0)>\alpha(t^e=0)$. As shown, α is strictly increasing in t^e for $1>\alpha>0$.

Note also that the wage differential is unchanged as a result of this anticipated tariff ($W^y=\lambda W^x$ at both A and B). Only if the expected tariff is greater than $\lambda/(1-\varepsilon)-1$ (in which case $W^y>\lambda W^x$ at $\alpha=1$), or if the actual tariff is partially unanticipated, will the private sector fail to fully offset the impact of the tariff on the resulting wage differential, and hence on the income distribution. The effect of an unexpected tariff is shown by point C in the diagram. Here the labor allocation is unaffected by the imposition of (surprise) protection, and the wage differential is reduced by the full amount of the tariff.

This feature of the model is depicted in Figure 2 where, using the lemma, the complete set of possible equilibrium combinations of wage differentials and fully anticipated tariffs is depicted by the locus abc. As reflected in the line segment ab, for anticipated tariffs satisfying $\lambda/(1-\varepsilon)>1+t\geq 1$, the relative wage is left unaffected at $W^y/W^x=\lambda$. For $1/(1-\varepsilon)\geq 1+t\geq\lambda/(1-\varepsilon)$, all workers remain in their pre-shock sectors ($\alpha=1$), and (13) implies a linear relationship between W^y/W^x and $1+t$ with positive slope $(1-\varepsilon)$, reflected in the line segment bc. At $1+t=1/(1-\varepsilon)$, wages in the two sectors are equalized, and higher tariffs need not be considered.

Also shown in Figure 2 are combinations of W^y/W^x and $1+t$ that would be attainable if the tariff were *un*anticipated. Using (2) and (5), the relative wage in the two sectors can be written as a function of both the actual and the expected tariffs as

$$(14)\quad \frac{W^y}{W^x}=\frac{(1-\varepsilon)(1+t)f'(\alpha(t^e))}{f'([(1-\alpha(t^e))\lambda+1])}.$$

For any (fixed) expected tariff level t^e, the intersectoral allocation of labor is given (α is fixed), and (14) describes a line with positive slope which crosses the equilibrium locus abc where the actual tariff equals the expected ($t^e=t$). Thus, for example, the dashed line labeled $t^e=0$ represents combinations of W^y/W^x and $1+t$ attainable if workers expect the government to maintain free trade after the terms-of-trade shock. This line

crosses the locus *abc* at $t=0$ (where actual protection is equal to expected) and has a positive slope which can be derived from the right-hand side of (14) for $t^e=0$.[5]

Figure 2 reiterates the point that, when workers expect tariff levels below the critical value of $\lambda/(1-\varepsilon)-1$, only unanticipated protection can alter the wage differential from its equilibrium value of λ. This is the sense in which, along *ab*, a government that wishes to redistribute income from low- to high-marginal utility-of-income workers has an incentive to provide unexpected protection.[6]

The equilibrium must lie on the *abc* locus of Figure 2, since any point not on this locus would involve some unexpected protection. Which point on the *abc* locus is chosen depends in part on the government's ability to influence workers' expectations. For example, the combination of W^y/W^x and $1+t$ represented by the point z in Figure 2 can only be an equilibrium if the government can impose the tariff t_1 and at the same time induce the domestic labor force to expect this level of protection. Any incentive to provide surprise protection at a point such as z will tend to undermine the credibility of the policy announcement, and may preclude z from being a feasible equilibrium point available to the government. We will return to these issues in the next two sections.

III. The Optimal Tariff Policy

In this section we characterize the optimal tariff, under the assumption that the government can undertake a binding commitment to a particular trade policy, and thus can influence the expectations of private agents concerning its behavior. With reference to Figure 2, the ability to undertake binding commitments allows the government to choose any point on the *abc* locus.

Using the notation introduced in Section I, the first-order conditions for the government problem are

$$(15)\quad \frac{\partial J}{\partial t}=\frac{\partial \alpha}{\partial t}(V^y-V^{yx})+(1-\varepsilon)\cdot \times\left[V_p^x+\alpha V_p^y+(1-\alpha)V_p^{yx}\right] +V_I^x\left(\frac{\partial \phi^x}{\partial t}I+\phi^x\frac{\partial I}{\partial t}\right) +\alpha V_I^y\left(\frac{\partial \phi^y}{\partial t}I+\phi^y\frac{\partial I}{\partial t}\right) +(1-\alpha)V_I^{yx}\left(\frac{\partial \phi^{yx}}{\partial t}I+\phi^{yx}\frac{\partial I}{\partial t}\right)=0.$$

In equilibrium (and hence with no unanticipated protection), any worker originally in sector y must be indifferent between staying or moving to sector x. Consequently, $V^y=V^{yx}$, $V_I^y=V_I^{yx}$, and $\phi^y=\phi^{yx}$. Making use of (11) to eliminate V_p^x, V_p^y, and V_p^{yx}, equation (15) can be rewritten as

$$(16)\quad \frac{\partial J}{\partial t}=I\left[V_I^x\frac{\partial \phi^x}{\partial t}+\alpha V_I^y\frac{\partial \phi^y}{\partial t} +(1-\alpha)V_I^y\frac{\partial \phi^{yx}}{\partial t}\right] -\left[\phi^xV_I^x+\phi^yV_I^y\right]\left[(1-\varepsilon)C^y-\frac{\partial I}{\partial t}\right]=0.$$

The first term on the right-hand side of (16) represents the marginal benefit of the tariff, in the form of income redistribution from high to low income workers. The second term represents the marginal cost of the tariff, net of tariff revenues, coming in the form of distortions in both consumption and production. At the optimum, the marginal cost

[5]Since α is increasing in t^e over the range $\lambda/(1-\varepsilon)>1+t^e\geq 1$, it follows from (14) that the slope of the dashed lines in Figure 2 is decreasing in t^e over this range. For $1+t^e\geq\lambda/(1-\varepsilon)$, we have from the lemma that $\alpha=1$. In this case, the slope of the dashed lines simplifies to $(1-\varepsilon)$. Therefore, for $1+t^e\geq\lambda/(1-\varepsilon)$, the dashed line coincides with the portion *bc* of the equilibrium locus *abc*.

[6]If the expected tariff is greater than or equal to $\lambda/(1-\varepsilon)-1$, the pre-shock intersectoral allocation of labor will be maintained ($\alpha=1$), and whether the actual tariff level is anticipated or not has no bearing on its effectiveness in redistributing income between workers in the two sectors. Thus, the incentive to surprise is not present along *bc*.

ciated with the political economy literature. Rather, it is a consequence of the government's inability to precommit to trade policies that, *ex post*, it would not find optimal to pursue.

Moreover, the policy requirement of time consistency can lead to a reversal of the traditional normative ordering of tariffs and subsidies as instruments of trade policy: under certain conditions on the parameter values of the model, we show that a time-consistent tariff policy is preferred to a time-consistent production subsidy. This result can contribute to an explanation of the empirical puzzle that was noted above: that is, why protection might take the form of trade distortions rather than of production or consumption subsidies.

Finally, these theoretical results contain a clear normative implication for improving on the time-consistent but suboptimal equilibrium: the government should be enabled to undertake binding commitments concerning its future behavior. From an operational point of view, this is suggestive of the important role that could be performed by an international organization like the GATT: namely, to enforce the domestic commitments to a policy of free trade. The GATT was originally conceived to facilitate international cooperation among individual countries; the results of the paper suggest that this institution can—and presumably to some extent already does—perform an equally crucial role in enforcing the cooperative outcome in a setting in which the strategic interaction is between each country and its own domestic residents.

The remainder of the paper proceeds as follows. Section I presents the model within which our analysis will be carried out. Section II considers the role of expected protection in determining the relocation of labor across sectors. The optimal tariff policy under the assumption that the government can undertake a binding commitment is derived in Section III. The time-consistent policy is derived in Section IV, and the results are compared with those of the previous section. Section V explores conditions under which a time-consistent tariff policy would be preferred to a time-consistent production subsidy. A concluding discussion appears in Section VI.

I. The Model

We consider a small open economy with two traded goods and two inputs to production, labor and capital. Capital is immobile across sectors, and technology is homogeneous of degree one in both inputs. To simplify notation, we assume that both sectors share the same production function, and that each sector is endowed with one unit of capital. The production technologies are given by

$$(1)\qquad i = f(N^i), \qquad i = x, y,$$
$$f'(\cdot) > 0, \quad f''(\cdot) < 0,$$

where $N^i \equiv$ labor employed in sector i; $x \equiv$ exported good; and $y \equiv$ imported good.

In each sector, firms combine labor with their fixed stock of capital, up to the point where the value of labor's marginal product is equated to the nominal wage measured in terms of any numeraire, W^i:

$$(2)\qquad p^x f'(N^x) = W^x,$$
$$p^y(1+t)f'(N^y) = W^y,$$

where p^x is the world price of the exported good, t is the *ad valorem* tariff on imports, and p^y is the world price of imports. Wages are assumed to be perfectly flexible so that (2) yields the nominal wage that clears the labor market in each sector.

Throughout the paper, we will consider the reaction of private agents and of the government to a terms-of-trade shock that lowers the world price of good y by an amount ε. The case of a negative shock to the price of exports is symmetric, except that the government would choose an export subsidy rather than a tariff. Prior to the realization of ε: (i) labor is assumed to be allocated equally between the two sectors, with the initial sectoral employment normalized to unity; (ii) goods units are chosen so that the world prices of x and of y are equal, and their common prices normalized to unity; and (iii) free trade is assumed to prevail in the domestic country. Given this,

equation (2) implies that, prior to the terms-of-trade shock, nominal wages are equal in the two sectors: $W^x = W^y$.

The aggregate supply of labor in the economy as a whole is assumed fixed. Labor is mobile between sectors, and reallocates in response to the terms-of-trade shock on the basis of the expected intersectoral wage differential, W^{ye}/W^{xe}. A central feature of the model is that, in the absence of binding commitments, the government cannot irrevocably set tariff policy before the labor reallocation decision is made. In other words, the timing of the decisions after the shock is observed is that, either: (a) first workers reallocate and then a tariff is imposed; or (b) the labor reallocation and the tariff decisions are made simultaneously. Since the private sector is atomistic and takes the government tariff as given, there is no relevant distinction between (a) and (b): both are valid interpretations of the equilibrium presented below. Under either of these timing assumptions, and using (2), the expected wage differential subsequent to the shock is given by

$$\text{(3)} \quad \frac{W^{ye}}{W^{xe}} = \frac{(1-\varepsilon)(1+t^e)f'(N^y(t^e))}{f'(N^x(t^e))}.$$

Equation (3) makes clear that the expected wage differential depends on the expected tariff, t^e, since the actual tariff is observed only once the reallocation is completed. The relationship between W^{ye}/W^{xe} and t^e will be analyzed more thoroughly in the next section.

A crucial assumption of the model is that labor's intersectoral mobility comes only at a cost. Specifically, we assume that, whenever one unit of labor moves from one sector to the other, its marginal product falls by the fraction $(1-\lambda)$, $1>\lambda>0$. This hypothesis can be motivated, for instance, by the notion that each worker has to acquire some sector-specific human capital, through experience or through training, before it can become as productive as workers already employed in that sector. Thus one can interpret $(1-\lambda)$ as the fraction of labor time that a worker relocating from one sector to the other has to spend in acquiring new productive skills.[2] Since we consider a one-period model, the issue of how the marginal product of the newly hired workers evolves over time does not arise. Section VI briefly discusses how to extend this approach to an explicit intertemporal framework.

From this assumption it follows that in equilibrium the expected wage differential between the two sectors of the economy is constrained by the following inequalities,

$$\text{(4)} \quad \frac{1}{\lambda} \geq \frac{W^{ye}}{W^{xe}} \geq \lambda.$$

If either inequality is violated, workers would find it optimal to move from one sector to the other, until condition (4) is satisfied. We abstract throughout the paper from corner solutions in which the entire labor force locates in a single sector.

The subsequent analysis will be simplified by expressing labor in efficiency units. Thus, letting α denote the fraction of labor that remains in sector y after all reallocation has taken place, the effective quantity of labor employed in each sector when the adjustment to the shock is completed is given by

$$\text{(5)} \quad N^y = \alpha, \quad N^x = (1-\alpha)\lambda + 1.$$

The first equation in (5) follows from having normalized to unity the quantity of labor initially in each sector. The second equation incorporates the notion that N^x is expressed

[2] See Gary Becker, 1962, and Baldwin, 1984b, for a more detailed motivation of this assumption. Baldwin, 1982, contains a discussion of the incomplete nature of adjustment assistance programs in eliminating the private costs of intersectoral labor movements. Michael Mussa, 1982, and Joshua Aizenman and Jacob Frenkel, 1986, model imperfectly mobile labor in a more general way, allowing for a continuum of λ's that reflect individual-specific moving costs. Modeling imperfectly mobile labor in this way would alter the characterization of the optimal policy in our model, but would leave unaffected our basic result that the opportunity for discretion in trade policy leads to excessive protection. Finally, the formal analysis would be different, but the qualitative results would remain unchanged, if the mobility costs were modeled as subjective disutility costs rather than output losses, or as payments made to a third sector responsible for moving workers between sectors.

in efficiency units, and that labor which has relocated to sector x from y has a marginal product equal only to a fraction λ of the marginal product of labor already employed in x. The fraction α of labor that remains in sector y is determined endogenously in response to the expected wage differential. The determination of α is the focus of the next section.

Define I as national disposable income valued at domestic prices. Imposing the condition of balanced trade at world prices, abstracting from domestic taxes, and assuming that the tariff is nonprohibitive, it follows that, subsequent to the shock

$$(6)\quad I=f(N^x)+(1-\varepsilon)(1+t)f(N^y)+T,$$

where the tariff revenue T is defined by

$$(7)\qquad T=t(C^y-f(N^y))(1-\varepsilon),$$

C^y being aggregate demand for the imported good y. Substituting (7) into (6), national income valued at domestic prices is given by

$$(8)\qquad I=f(N^x)+(1-\varepsilon)f(N^y)+t(1-\varepsilon)C^y.$$

In order to focus on the redistributive consequences of tariffs for the labor allocation decision, we assume that the distribution of income is determined solely on the basis of the wage differential between the two sectors. Thus, income from capital and tariff revenues is distributed to each worker in proportion to the share of his labor income in the economywide wage bill.[3] Define the income share variable, ϕ, as

$$(9)\quad \phi^i\equiv\frac{I^i}{I}=\frac{W^i}{W^x+\alpha W^y+(1-\alpha)W^{yx}},\qquad i=x,y,yx,$$

where I^i is total disposable income of a worker of the ith type (valued at domestic prices), and the superscripts x, y, and yx denote workers originally in sector x who remain there, workers originally in sector y who remain there, and workers originally in sector y who move to sector x, respectively.

Each worker consumes a bundle of x and y, chosen so as to maximize an identical homothetic utility function, subject to a standard budget constraint. The indirect utility function of a representative consumer of the ith type is assumed to exhibit diminishing marginal utility of income, and can be written in terms of the previous notation as

$$(10)\quad V^i=V(p^x,p^y,I^i),\quad i=x,y,yx.$$

Letting V_p^i and V_I^i denote the partial derivatives of (10) with respect to p^y and I^i, respectively, the consumption of y on the part of consumers of the ith type can be expressed, using Roy's identity, as

$$(11)\quad C_i^y=-\frac{V_p^i}{V_I^i}=\phi^iC^y,\quad i=x,y,yx.$$

The second equality follows from the assumption that the common utility function is homothetic.

Finally, the government chooses a level of protection in response to the shock ε, so as to maximize a welfare function defined over the indirect utility functions of the three types of workers,

$$(12)\qquad J=\alpha V^y+(1-\alpha)V^{yx}+V^x.$$

Thus, the weights in the social welfare function given to the three types of workers are chosen to be proportional to the size of the group to which each type belongs. A property of the model will be that, in equi-

[3] Hence, we are assuming that lump-sum instruments are available to distribute tariff revenue to consumers, but that these instruments are insufficiently flexible to achieve the optimal distribution of income across workers. In the extreme, as assumed here, the distribution of tariff revenues leaves the income distribution completely unaffected. If tariff revenues could be redistributed optimally, none of the qualitative results of the paper would change, except that the optimal tariff policy would in general involve some positive level of protection.

librium, $V^y = V^{yx}$ (that is, workers must be indifferent between staying in y or moving to x). Hence, the coefficient α in (12) affects the equilibrium value of J only indirectly, through its effect on the equilibrium trade policy, and not directly as a weighting factor. We mention the implications of more or less egalitarian weighting factors for the resulting trade policy in a later section.

We assume the absence of any market mechanisms for reallocating the risk associated with the terms-of-trade shock. If such private insurance markets existed and worked perfectly, there would be no role for government intervention in the form of protection in our model. Support for such an assumption is contained in Eaton and Grossman (1985).

II. The Reallocation of Labor Across Sectors

As discussed in the previous section, workers will respond to a fall in the world price of imports by reallocating across sectors in order to take advantage of any expected wage differential in excess of λ. We assume that the world price of y falls by an amount $\varepsilon > 1-\lambda$. This inequality assures that, if a zero tariff were expected by workers in sector y (that is, if $t^e = 0$), some sectoral movement of labor from y to x would take place.[4]

The movement of workers across sectors in response to ε will assure that, in equilibrium, the expected wage differential subsequent to the shock will satisfy the inequalities in (4). Making use of equations (3) and (5), the expected wage differential can be rewritten as

$$\text{(13)} \qquad \frac{W^{ye}}{W^{xe}} = \frac{(1+t^e)(1-\varepsilon)f'(\alpha)}{f'([(1-\alpha)\lambda+1])}.$$

The fraction α of workers that remain in sector y after the shock is a function of t^e, defined implicitly by (13) and (4). The properties of α as a function of t^e are summarized in the following:

LEMMA: (*i*) *For* $1/(1-\varepsilon) \geq 1+t^e \geq \lambda/(1-\varepsilon)$, α *equals unity, and* W^{ye}/W^{xe} *is a continuous, differentiable, and strictly increasing function of* t^e, *with* $1 \geq W^{ye}/W^{xe} \geq \lambda$. (*ii*) *For* $\lambda/(1-\varepsilon) > 1+t^e \geq 1$, W^{ye}/W^{xe} *equals* λ, *and* α *is a continuous, differentiable, and strictly increasing function of* t^e, *with* $1 > \alpha > 0$.

PROOF:

For $\alpha = 1$, the expected wage differential W^{ye}/W^{xe} is determined by (13) as $W^{ye}/W^{xe} = (1+t^e)(1-\varepsilon)$. Therefore, if $1/(1-\varepsilon) \geq 1+t^e \geq \lambda/(1-\varepsilon)$ and $\alpha = 1$, equilibrium condition (4) is satisfied; given the initial allocation of labor, $W^{ye} \geq \lambda W^{xe}$. Consequently, when the condition in part (*i*) of the lemma is met, no worker finds it worthwhile to relocate from sector y to sector x. Finally, with $\alpha = 1$, the rest of part (*i*) follows immediately from (13). Alternatively, if $\lambda/(1-\varepsilon) > 1+t^e \geq 1$, then the second inequality in (4) will be violated if $\alpha = 1$; under the initial allocation of labor, $W^{ye} < \lambda W^{xe}$. Consequently, when the condition in part (*ii*) of the lemma holds, the second inequality in (4) will in equilibrium hold with equality for some $1 > \alpha > 0$, and labor will move from sector y to sector x until $W^{ye} = \lambda W^{xe}$. The rest of part (*ii*) then follows immediately by applying the implicit function theorem.

The lemma is illustrated in Figure 1. The horizontal axis measures the fraction α of workers remaining in sector y. The vertical axis measures the wage rate. Given the concavity of the production function, W^y is decreasing in α, and is represented for $t = 0$ by the downward sloping solid curve. Conversely, the wage that can be earned by moving to sector x, λW^x, is given by the upward sloping curve. If protection is neither anticipated nor forthcoming, the equilibrium allocation is given by $\alpha(t^e = 0)$, and corresponds to the point where the cost of reallocating is just equal to the wage differential, $W^y = \lambda W^x$. The imposition of a tariff shifts the W^y curve to the right, say to the dotted

[4] This can be seen by noting that, in the absence of any intersectoral labor movement, $N^x = N^y = 1$, and (3) implies $W^{ye}/W^{xe} = (1-\varepsilon) < \lambda$, which violates the labor market equilibrium condition (4).

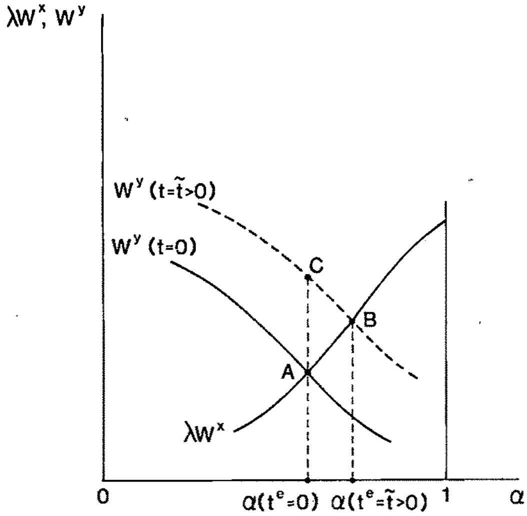


FIGURE 1

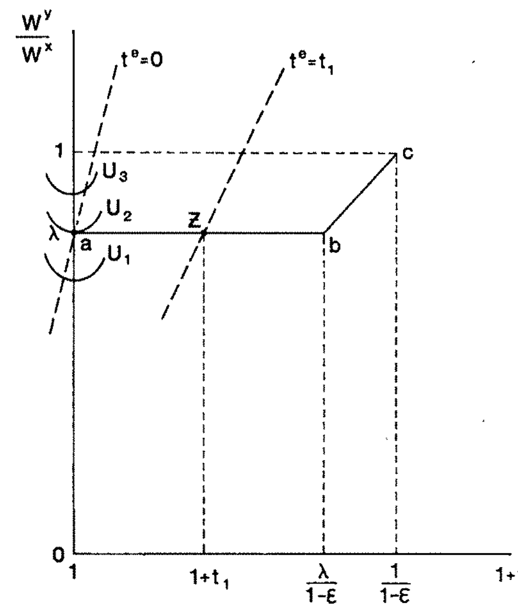


FIGURE 2

line of Figure 1. With the tariff fully anticipated, the equilibrium labor allocation is now given by point B in the diagram, where fewer workers have reallocated from sector y to x, that is, $\alpha(t^e=\tilde{t}>0)>\alpha(t^e=0)$. As shown, α is strictly increasing in t^e for $1>\alpha>0$.

Note also that the wage differential is unchanged as a result of this anticipated tariff ($W^y=\lambda W^x$ at both A and B). Only if the expected tariff is greater than $\lambda/(1-\varepsilon)-1$ (in which case $W^y>\lambda W^x$ at $\alpha=1$), or if the actual tariff is partially unanticipated, will the private sector fail to fully offset the impact of the tariff on the resulting wage differential, and hence on the income distribution. The effect of an unexpected tariff is shown by point C in the diagram. Here the labor allocation is unaffected by the imposition of (surprise) protection, and the wage differential is reduced by the full amount of the tariff.

This feature of the model is depicted in Figure 2 where, using the lemma, the complete set of possible equilibrium combinations of wage differentials and fully anticipated tariffs is depicted by the locus abc. As reflected in the line segment ab, for anticipated tariffs satisfying $\lambda/(1-\varepsilon)>1+t\geq 1$, the relative wage is left unaffected at $W^y/W^x=\lambda$. For $1/(1-\varepsilon)\geq 1+t\geq\lambda/(1-\varepsilon)$, all workers remain in their pre-shock sectors ($\alpha=1$), and (13) implies a linear relationship between W^y/W^x and $1+t$ with positive slope $(1-\varepsilon)$, reflected in the line segment bc. At $1+t=1/(1-\varepsilon)$, wages in the two sectors are equalized, and higher tariffs need not be considered.

Also shown in Figure 2 are combinations of W^y/W^x and $1+t$ that would be attainable if the tariff were *un*anticipated. Using (2) and (5), the relative wage in the two sectors can be written as a function of both the actual and the expected tariffs as

$$(14)\quad \frac{W^y}{W^x}=\frac{(1-\varepsilon)(1+t)f'(\alpha(t^e))}{f'([(1-\alpha(t^e))\lambda+1])}.$$

For any (fixed) expected tariff level t^e, the intersectoral allocation of labor is given (α is fixed), and (14) describes a line with positive slope which crosses the equilibrium locus abc where the actual tariff equals the expected ($t^e=t$). Thus, for example, the dashed line labeled $t^e=0$ represents combinations of W^y/W^x and $1+t$ attainable if workers expect the government to maintain free trade after the terms-of-trade shock. This line

crosses the locus *abc* at $t=0$ (where actual protection is equal to expected) and has a positive slope which can be derived from the right-hand side of (14) for $t^e=0$.[5]

Figure 2 reiterates the point that, when workers expect tariff levels below the critical value of $\lambda/(1-\varepsilon)-1$, only unanticipated protection can alter the wage differential from its equilibrium value of λ. This is the sense in which, along *ab*, a government that wishes to redistribute income from low- to high-marginal utility-of-income workers has an incentive to provide unexpected protection.[6]

The equilibrium must lie on the *abc* locus of Figure 2, since any point not on this locus would involve some unexpected protection. Which point on the *abc* locus is chosen depends in part on the government's ability to influence workers' expectations. For example, the combination of W^y/W^x and $1+t$ represented by the point z in Figure 2 can only be an equilibrium if the government can impose the tariff t_1 and at the same time induce the domestic labor force to expect this level of protection. Any incentive to provide surprise protection at a point such as z will tend to undermine the credibility of the policy announcement, and may preclude z from being a feasible equilibrium point available to the government. We will return to these issues in the next two sections.

III. The Optimal Tariff Policy

In this section we characterize the optimal tariff, under the assumption that the government can undertake a binding commitment to a particular trade policy, and thus can influence the expectations of private agents concerning its behavior. With reference to Figure 2, the ability to undertake binding commitments allows the government to choose any point on the *abc* locus.

Using the notation introduced in Section I, the first-order conditions for the government problem are

$$(15)\quad \frac{\partial J}{\partial t}=\frac{\partial \alpha}{\partial t}(V^y-V^{yx})+(1-\varepsilon)\cdot \times\left[V_p^x+\alpha V_p^y+(1-\alpha)V_p^{yx}\right] + V_I^x\left(\frac{\partial \phi^x}{\partial t}I+\phi^x\frac{\partial I}{\partial t}\right) + \alpha V_I^y\left(\frac{\partial \phi^y}{\partial t}I+\phi^y\frac{\partial I}{\partial t}\right) +(1-\alpha)V_I^{yx}\left(\frac{\partial \phi^{yx}}{\partial t}I+\phi^{yx}\frac{\partial I}{\partial t}\right)=0.$$

In equilibrium (and hence with no unanticipated protection), any worker originally in sector y must be indifferent between staying or moving to sector x. Consequently, $V^y=V^{yx}$, $V_I^y=V_I^{yx}$, and $\phi^y=\phi^{yx}$. Making use of (11) to eliminate V_p^x, V_p^y, and V_p^{yx}, equation (15) can be rewritten as

$$(16)\quad \frac{\partial J}{\partial t}=I\left[V_I^x\frac{\partial \phi^x}{\partial t}+\alpha V_I^y\frac{\partial \phi^y}{\partial t} +(1-\alpha)V_I^y\frac{\partial \phi^{yx}}{\partial t}\right] -\left[\phi^x V_I^x+\phi^y V_I^y\right]\left[(1-\varepsilon)C^y-\frac{\partial I}{\partial t}\right]=0.$$

The first term on the right-hand side of (16) represents the marginal benefit of the tariff, in the form of income redistribution from high to low income workers. The second term represents the marginal cost of the tariff, net of tariff revenues, coming in the form of distortions in both consumption and production. At the optimum, the marginal cost

[5]Since α is increasing in t^e over the range $\lambda/(1-\varepsilon)>1+t^e\geq 1$, it follows from (14) that the slope of the dashed lines in Figure 2 is decreasing in t^e over this range. For $1+t^e\geq\lambda/(1-\varepsilon)$, we have from the lemma that $\alpha=1$. In this case, the slope of the dashed lines simplifies to $(1-\varepsilon)$. Therefore, for $1+t^e\geq\lambda/(1-\varepsilon)$, the dashed line coincides with the portion *bc* of the equilibrium locus *abc*.

[6]If the expected tariff is greater than or equal to $\lambda/(1-\varepsilon)-1$, the pre-shock intersectoral allocation of labor will be maintained ($\alpha=1$), and whether the actual tariff level is anticipated or not has no bearing on its effectiveness in redistributing income between workers in the two sectors. Thus, the incentive to surprise is not present along *bc*.

and benefit of the tariff must be equated. We can now characterize the optimum tariff policy as follows:

PROPOSITION I: *The optimal tariff policy is either free trade or the imposition of a sufficiently high tariff $\bar{t}$ that prevents any sectoral reallocation of labor from taking place.*

PROOF:

To prove this proposition, we need only show that $t=0$ is a solution to (16) and that (16) is violated for any t satisfying $\lambda/(1-\varepsilon)>1+t>1$, that is, for any positive tariff consistent with $\alpha<1$.[7] For $\lambda/(1-\varepsilon)>1+t\geq 1$ and $t^e=t$, the terms $\partial\phi^x/\partial t|_{t^e=t}$, $\partial\phi^y/\partial t|_{t^e=t}$, and $\partial\phi^{yx}/\partial t|_{t^e=t}$ are all zero, implying the absence of any marginal redistributive gains and a zero marginal benefit (the first term in (16)) from a tariff over this range. At the same time, (2), (5), and (8) imply that

$$\text{(17)}\quad \left.\frac{\partial I}{\partial t}\right|_{t^e=t} = -\left[W^x\lambda - \frac{W^y}{1+t}\right]\left.\frac{\partial \alpha}{\partial t}\right|_{t^e=t} + (1-\varepsilon)C^y + t(1-\varepsilon)\left.\frac{\partial C^y}{\partial t}\right|_{t^e=t}$$

so that $\left.\dfrac{\partial I}{\partial t}\right|_{t^e=t} = (1-\varepsilon)C^y$ for $t=0$,

and $\left.\dfrac{\partial I}{\partial t}\right|_{t^e=t} < (1-\varepsilon)C^y$ for $t>0$.

The marginal cost of the tariff (the second term in (16)) is therefore zero at $t=0$ but strictly positive for $t>0$. As such, (16) is satisfied for $t=0$ but violated for any t such that $\lambda/(1-\varepsilon)>1+t>1$.

Consequently, the optimal tariff policy when the government can make a binding commitment will be either one of free trade, or the choice of a tariff $\bar{t}$ no smaller than $\lambda/(1-\varepsilon)-1$. In the latter case, no workers will choose to relocate from sector y to x.

Figure 2 depicts the choice of the optimal tariff policy when the government can precommit. Recall that *abc* represents the locus of possible equilibrium combinations of relative wage and fully anticipated tariff. Three government indifference curves, defined over the relative wage W^y/W^x and fully anticipated tariff $(1+t)$, are labeled U_1, U_2, and U_3. The government's bliss point is $(1,1)$, where wages are completely equalized and there are no trade distortions. For relative wages below unity, the government's objective function is strictly increasing in W^y/W^x. As noted above, the marginal cost of a small increment in the tariff is zero at $t=0$ and strictly positive for $t>0$. As such, the slope of any government indifference curve in $(W^y/W^x, 1+t)$ space is zero at $t=0$ and strictly positive for $t>0$. Point *a* in Figure 2 represents the situation in which free trade is associated with the highest attainable indifference curve on the equilibrium *abc* locus, and hence is the optimal policy.

Of course, this is not the only possibility. In the case where the optimal policy is the imposition of a tariff that prevents any sectoral reallocation of labor from occurring, the equilibrium will be represented by a tangency point on the *bc* portion on the *abc* locus in Figure 2.[8] As noted in the previous section, along *bc* there are no gains from providing surprise protection, and so the government has no incentive to act unexpectedly. As such, when the optimal policy is one which prevents any reallocation of labor from taking place, the issue of credibility that is the focus of this paper does not arise. Consequently, in the remaining sections we concentrate on the case in which free trade is the optimal policy. It is easily shown that, for any $0<\varepsilon<1$, free trade must be the optimal policy if the cost of relocating is not "too large," that is, if λ is close enough to one.

[7]The second-order conditions will be met at $t=0$. For $1/(1-\varepsilon)>1+t>1$, we assume that the second-order conditions hold. Relaxing this assumption would complicate but not alter the nature of our results.

[8]If the optimal tariff is strictly positive, it is characterized by the following expression, provided that the second-order conditions are satisfied:

$$\bar{t} = \frac{I\left[V^x_{\bar{t}}/V^y_{\bar{t}} - 1\right]}{\left[1+(1-\varepsilon)(1+\bar{t})\right]\cdot\left[V^x_{\bar{t}}/V^y_{\bar{t}} + (1-\varepsilon)(1+\bar{t})\right]\partial C^y/\partial t|_{t^e=\bar{t}}}.$$

IV. The Time Inconsistency of Free Trade

In this section we consider an alternative institutional setup, in which the government is unable to precommit to a specific course of action. The kind of discretion embodied, for example, in the escape clause (section 201 of the Trade Act of 1974) could make it impossible for a government to credibly commit to any trade policy which, *ex post*, was not optimal to pursue. Here, the government cannot influence worker expectations by pre-announcing a tariff policy, since it has no credible mechanism with which to bind itself to the announced intention. Hence, the government must take the expected tariff t^e as given when it optimizes with respect to the actual tariff t. Since α, the fraction of labor that remains in sector y, depends on t^e but not on t (see the lemma in Section II), this is equivalent to saying that the government takes the allocation of labor, and hence the production side of the economy, as given when computing the optimal tariff. The solution to this problem provides the time-consistent equilibrium or, more precisely, the subgame perfect Nash equilibrium of a game between the government and the labor market.

Consider first the case in which the cost of relocating is not "too large," so that the optimal policy considered in the previous section is one of free trade. To determine the time-consistent tariff policy in this case, note that the first-order conditions of the government's problem are still given by equation (16) in Section III. However, the terms $\partial I/\partial t$ and $\partial\phi^i/\partial t$ that appear in (16) are now different, because of the requirement that t^e and thus α be taken as given. Specifically, the effect of the tariff on national income with t^e fixed at t_0^e can be derived from (8) as

$$(18)\quad \left.\frac{\partial I}{\partial t}\right|_{t^e=t_0^e}=(1-\varepsilon)C^y + t(1-\varepsilon)\left.\frac{\partial C^y}{\partial t}\right|_{t^e=t_0^e}.$$

Thus, the marginal cost of a small unexpected increase in the tariff—the second term in (16)—is zero at $t=t_0^e=0$, since $\partial I/\partial t|_{t^e=t_0^e}=(1-\varepsilon)C^y$, and is positive for any $t=t_0^e>0$, since then $\partial I/\partial t|_{t^e=t_0^e}<(1-\varepsilon)C^y$.

The effect of the tariff on the distribution of income, for a fixed t^e, can be derived from (2) and (9). After some algebraic computations we obtain

$$(19)\quad \left.\frac{\partial\phi^x}{\partial t}\right|_{t^e=t_0^e}=\frac{-\alpha(t_0^e)\lambda}{(1+\lambda)^2(1+t)}<0;$$
$$\left.\frac{\partial\phi^{yx}}{\partial t}\right|_{t^e=t_0^e}=\lambda\left.\frac{\partial\phi^x}{\partial t}\right|_{t^e=t_0^e}<0;$$
$$\left.\frac{\partial\phi^y}{\partial t}\right|_{t^e=t_0^e}=\lambda\frac{(1+(1-\alpha(t_0^e))\lambda)}{(1+\lambda)^2(1+t)}>0.$$

Therefore, using (19), the marginal benefit of unexpected protection (the first term in (16)) at $t=t_0^e=0$ is given by $I[V_I^y-V_I^x][\alpha(t_0^e)\lambda]/[(1+\lambda)^2]$, which is strictly positive: with t^e held at zero, the marginal benefit of an unexpectedly higher tariff exceeds its marginal cost. Hence, the government has an incentive to surprise the domestic labor force with a strictly positive level of protection. The time-consistent tariff is determined by the dual conditions that the tariff is fully expected and that the marginal gain from unexpected protection is equal to the marginal cost. Plugging (18) and (19) into (16), such a condition simplifies to the following expression, which characterizes the time-consistent tariff $\hat{t}$ as

$$(20)\quad \hat{t}(1+\hat{t}) = \frac{\alpha\lambda I\left[V_I^x/V_I^y-1\right]}{(1+\lambda)(1-\varepsilon)(V_I^x/V_I^y+\lambda)\,\partial C^y/\partial t|_{t^e=\hat{t}}} > 0.$$

The right-hand side of (20) is strictly positive under the assumptions of the model. The left-hand side of (20) must therefore be strictly positive as well, which means that $0<\hat{t}<\lambda/(1-\varepsilon)-1$.[9] Table 1 contains

[9]Negative values of $\hat{t}$ satisfying (20) can be ruled out since the negative tariff would have to be large enough in absolute value to make $(1+\hat{t})$, and thus the domestic

TABLE 1—TARIFFS AND SUBSIDIES COMPARED

	Computations[a]						
	$\hat{t}$	$\alpha(\hat{t})$	$J(\hat{t})$	$\hat{s}$	$J(\hat{s})$	$\alpha(\bar{t}=0)$	$J(\bar{t}=0)$
$\lambda=.75$	.3	.92	1.7561	1	1.7835	.78	1.7820
$\lambda=.9$	.08	.74	1.815	1	1.7835	.70	1.818

[a]The computations are based on the following assumptions. The utility function is $U=\frac{1}{2}\ln x+\frac{1}{2}\ln y$. The production function is $x=\ln N^x$; $y=\ln N^y$. The labor force in each sector prior to the realization of the shock has been normalized to equal 10. Finally, the size of the shock to the world price of y has been taken to be $\varepsilon=.5$. The symbols in the table are defined as follows: $\bar{t}\equiv$ optimal tariff; $\hat{t}\equiv$ time-consistent tariff; $\alpha\equiv$ fraction of the labor force remaining in sector y, as a function of the tariff; $\hat{s}\equiv$ time-consistent subsidy; $J(\ )\equiv$ value taken in equilibrium by the social welfare function, as a function of the various policy instruments; $\lambda\equiv$ ratio between the marginal product of the labor that has just moved to sector x from sector y, and the marginal product of the labor originally in sector x that remains there.

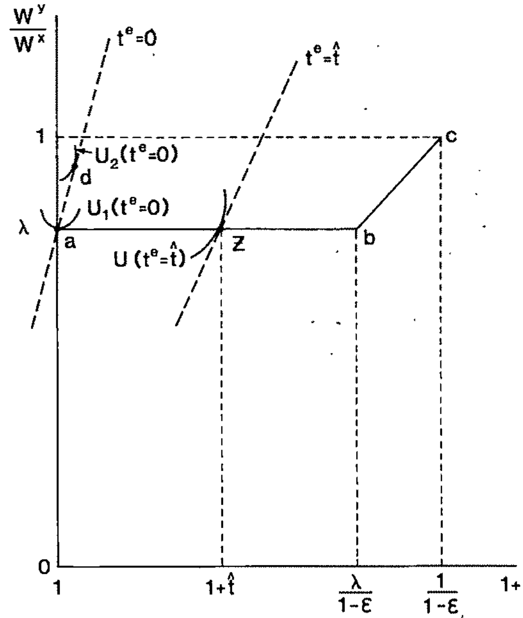


FIGURE 3

numerical values of $\hat{t}$ for some simple utility and production functions. These examples indicate that, for shocks to the world price of imports which are of sufficient magnitude, the time-consistent tariff is quite large, even though, for the same set of parameter values, the optimal policy is one of free trade.

Figure 3 illustrates the time-consistent solution. A family of government indifference curves defined over W^y/W^x and $(1+t)$ exists for each level of the expected tariff. Consider the indifference curve corresponding to $t^e=0$ and passing through point a on the equilibrium locus abc. Repeating the arguments of the previous section, and making use of (18) and (19), it can be shown that the government indifference curve is flat at $t=0$ and upward sloping for $t>0$ in a neighborhood of point a. However, with t^e fixed at zero, the locus of feasible points is now given by the upward sloping dashed line labeled as $t^e=0$: as noted, the government can increase the relative wage above λ by imposing an unexpected positive tariff. Hence, the government has an incentive to surprise workers by moving to point d in Figure 3, where a government indifference curve in the $t^e=0$ family is tangent to the $t^e=0$ locus.

Of course, since this level of protection is unexpected by the labor force, point d is not an equilibrium. The time-consistent equilibrium will occur at a point such as z on the abc locus in Figure 3, where the tariff is fully expected ($t^e=\hat{t}$) and an indifference curve in the family of curves for $t^e=\hat{t}$ is tangent to the $t^e=\hat{t}$ locus. A policy such as $\hat{t}$ is credible since, given that workers expect this level of protection, the government has no incentive to provide a level of protection other than $\hat{t}$. Since $0<\hat{t}<\lambda/(1-\varepsilon)-1$, the time-consistent equilibrium must lie on the horizontal segment ab, where $W^y/W^x=\lambda$. Hence, when free trade is optimal, the time-consistent equilibrium will involve a strictly positive level of protection but the same income distribution that would prevail with free trade. As such, social welfare is higher if precommitment to a policy of free trade is possible.

price of y, $(1-\varepsilon)(1+\hat{t})$, negative. As (20) applies to the case in which the optimal tariff is zero, values of $\hat{t}$ greater than $\lambda/(1-\varepsilon)-1$ can also be ruled out, since when greater than this critical value, the time-consistent tariff must correspond to the optimal tariff (see Section III).

It is also possible to show that the time-consistent tariff increases with the weight assigned by the government to the utility of the workers who remain in the injured sector. Thus, for instance, a more egalitarian government, which assigns a greater importance to the marginal redistributive gains of unexpected protection, will lower the welfare of domestic citizens, since the income distribution remains unaffected but the degree of distortionary protection is higher than with a government that is less motivated by redistributive objectives.[10]

Finally, recall that if the optimal policy is not free trade but rather the imposition of a tariff greater than the critical value $\lambda/(1-\varepsilon)-1$, then it is represented by a point on the line segment *bc*. Since there is no incentive to surprise along *bc*, such a policy is time consistent. Thus, the ability to make commitments becomes relevant in this model when $\hat{t}<\lambda/(1-\varepsilon)-1$, that is, precisely when free trade is the optimal policy. We summarize these results in the following:

PROPOSITION II: *When free trade is the optimal policy, it is not time consistent. The time-consistent policy involves a socially excessive level of protection.*

V. Tariffs vs. Production Subsidies

In this section we show that, when a government is constrained to the use of time-consistent policies, a tariff may be preferred to a production subsidy. As such, if the government has some choice concerning the policy instrument over which the game with workers will be played, but is unable to commit to a time-inconsistent strategy for the particular instrument chosen, it may rationally choose to operate in a regime in which only tariffs, and not subsidies, can be used.[11]

To demonstrate this result, we first show that the time-consistent subsidy is strictly higher than the time-consistent tariff. This occurs precisely because unexpected tariffs are, on the margin, more distortionary than unexpected subsidies and, as such, more costly to use. We then show that, if the social gains from a marginal redistribution of income are small enough, a time-consistent tariff policy will always be preferred to the imposition of a time-consistent subsidy.

As in the case of a tariff, an increase in the production subsidy imposes no distortionary costs on the production side of the economy as long as it is unexpected. But unlike a tariff, a production subsidy leaves consumption decisions completely undistorted as well. As such, given any expected subsidy level for which $W^y<W^x$, the marginal cost of providing an unexpectedly higher subsidy is zero while the marginal redistributive benefit is strictly positive. This implies that the time-consistent subsidy, $\hat{s}$, will be chosen at a level which equates W^y and W^x, since only then will the marginal costs and benefits of the subsidy be equated. Substituting s for t in (2), and setting equal the wage rates in the two sectors, we obtain the time-consistent

[10] This result, that in a time-consistent equilibrium society may be better off by appointing a government that does not reflect its true preferences, is not unique to this model—see for instance Kenneth Rogoff (1985).

[11] An important question concerns the mechanism by which a government could commit to the use of one policy tool over another, even if commitments to the intensity with which the policy tool is used are infeasible. While not modeled explicitly in this paper, a prior stage to our one-stage game could be introduced in which institutions are created to carry out the policies of the second stage. Here the government might consider the creation of an institution capable of carrying out trade policies, thus committing itself to use only trade restricting policies in the second period. Alternatively, it might choose to create an institution to impose domestic subsidies and taxes. Finally, the government may choose not to create any institution, thus committing itself to no intervention whatsoever in the second period. Since the choice of institution must be made prior to the realization of the shock, the possibility of intervention as the optimal response to a future shock will reduce the desirability of precommitting to inaction by following the third option. The result of Section V can be interpreted as a rationale for why a government might wish to pursue the first option and precommit, through the prior choice of policy institutions, to react to terms-of-trade shocks with tariffs rather than with domestic subsidies. In practice, international and U.S. trade law provide avenues for the relatively prompt imposition of trade restraints as compared to the imposition of purely domestic subsidies and taxes.

subsidy

$$1+\hat{s}=\frac{1}{1-\varepsilon}. \tag{21}$$

The time-consistent tariff, by contrast, will never be set at a level that is so high as to equalize wages in the two sectors. As discussed in the previous section, the marginal cost of an unexpected tariff is positive for $t>0$. Hence, in the time-consistent equilibrium, the marginal benefit of providing an unexpectedly higher tariff must also be positive, which in turn implies that $W^y<W^x$, and consequently that

$$1+\hat{t}<\frac{1}{1-\varepsilon}=1+\hat{s}. \tag{22}$$

Thus, the time-consistent subsidy is strictly higher than the time-consistent tariff. However, whether the government objective function achieves a higher value with a time-consistent subsidy or tariff depends on the parameters of the model. The subsidy brings about larger production distortions than the tariff, since in equilibrium it is set at a higher level. On the other hand, the tariff introduces consumption distortions that would be absent with the subsidy.

It can be shown for a given shock ε that, if the social benefits of income redistribution are small enough, so that the time-consistent tariff is sufficiently low, then the tariff always welfare dominates the subsidy. For instance, consider what happens when the cost of relocating, $1-\lambda$, approaches zero, so that λ approaches one. In the same way that the optimal tariff $\bar{t}$ was characterized in Section III, it can be shown that the optimal subsidy $\bar{s}$ will be either zero or sufficiently high to prevent the relocation of labor, and will always be zero if λ is close enough to one, that is, if the cost of relocation is sufficiently small.[12] Therefore, for λ sufficiently close to one, both $\bar{s}$ and $\bar{t}$ will be zero. Further, with $\bar{t}=0$, the time-consistent tariff $\hat{t}$ will be characterized by (20), and also approaches zero as λ approaches one.[13] As a result, welfare under the time-consistent tariff can be made arbitrarily close to welfare under the optimal subsidy by bringing λ arbitrarily close to one. But for any λ strictly less than one, the time-consistent subsidy is fixed, and is given by (21) as $\hat{s}=1/(1-\varepsilon)-1$. Since welfare under the optimal subsidy $\bar{s}$ is strictly greater than welfare under the time-consistent subsidy $\hat{s}$ whenever $\bar{s}\neq\hat{s}$, choosing λ so that $\hat{t}$ is arbitrarily close to zero (and thus to $\bar{s}$) will ensure that welfare under the time-consistent tariff is strictly greater than that under the time-consistent subsidy.

The numerical example reported in Table 1 illustrates this result. If the world price of imports falls by 50 percent ($\varepsilon=.5$), and if it is very costly to reallocate labor across sectors ($\lambda=.75$), the time-consistent production subsidy is 100 percent, and the time-consistent tariff is 30 percent. For these parameter values, the time-consistent subsidy welfare dominates the time-consistent tariff. However, if the cost of reallocating labor across sectors falls ($\lambda=.9$), so that the equilibrium wage differential falls accordingly, then the marginal gains from income redistribution become smaller. As a result, the time-consistent tariff is reduced to 8 percent, whereas the time-consistent subsidy remains at 100 percent. In this second case, the time-consistent tariff welfare dominates the time-consistent subsidy. Also, note that in both cases the policy of free trade gives a higher social welfare than the time-consistent tariff.

We can summarize the foregoing discussion in the following:[14]

[12] In fact, $\bar{s}$ will either be zero or equal to $1/(1-\varepsilon)-1$.

[13] This can be seen by noting that the imposition of $\hat{t}$ will involve some equilibrium labor relocation ($\alpha<1$) so that $W^y/W^x=\lambda$ (see the lemma in Section II). Thus, as λ goes to one the equilibrium wage differential W^y/W^x goes to one as well, which implies that V_I^x/V_I^y approaches one. From (20), $\hat{t}$ therefore approaches zero.

[14] In comparing the subsidy and the tariff, we neglected the issue of how to finance the subsidy. An appropriate combination of production tax in sector x and subsidy in sector y can avoid any problems of how to raise the needed revenue. Alternatively, a lump-sum income tax which is neutral with respect to the distribution of income, as in Mayer and Raymond Riezman

PROPOSITION III: *In the time-consistent equilibrium, the production subsidy is strictly higher than the tariff. If the social gains from redistribution are small enough, the tariff welfare dominates the subsidy.*

VI. Generalizations and Conclusions

The model presented in the previous sections can be generalized in several directions. None of these extensions would change the nature of the main result, namely that the time-consistent trade policy involves more protection than the optimal policy would dictate; however, the formal details of the argument would be different. This section contains a brief discussion of one of the potentially more interesting extensions.

In particular, the one-shot game between the policymaker and the labor market that was analyzed in the previous sections could be extended to an intertemporal framework. The solution would be identical to the one given above if the same static game were repeated any finite number of times. However, the nature of the game would change if we dropped the assumption that there is no "learning by doing" for the workers that change sectors, that is, if we allowed λ to increase over time. The policymaker would now face a dynamic optimization problem, and this would add further opportunities for time inconsistencies of the optimal policies (see Kydland and Prescott, 1977). If, instead, the same static game were repeated an infinite number of times, or if an element of asymmetric information were incorporated into the finitely repeated game, then the policymaker would face incentives to maintain or establish a reputation. As is well known from the macroeconomic literature, these incentives could reduce the difference between the optimal and time-consistent policies.

Finally, it should be emphasized that while we have modeled a situation in which a policy of free trade is optimal but not time consistent, the result that discretion in trade policy can be costly is not limited to the special assumptions of this model. Much of the recent literature on the theory of international trade has analyzed models with some kind of market imperfection, and a role for activist trade policy commonly emerges. However, the particular nature of the market imperfections motivating the activist policy in many of these models contains a second, more subtle, implication: the activist trade policy must be pursued with discretion and with flexibility, judging each situation on a case-by-case basis. The model explored in the previous sections points out that increasing the discretion and flexibility of the government decision process may be counterproductive. Many of the same market imperfections that motivate trade policy intervention can also generate time inconsistencies in the implementation of the optimal activist policies. Whenever this happens, a government pursuing a discretionary trade policy finds itself trapped in a suboptimal equilibrium. Thus, a commitment to a simple set of trading rules may often be superior to an activist but discretionary trade policy.

(1985), will leave the results of Proposition III unaffected. Mayer and Riezman also consider a progressive lump-sum income tax as a means of financing the production subsidy. To the extent that such a lump-sum redistributive mechanism can be used to redistribute income in our model, it would reduce the need to use tariffs or subsidies as risk-sharing devices, but would otherwise leave unaltered our conclusions concerning the normative ranking of these time-consistent policies.

REFERENCES

Aizenman, Joshua and Frenkel, Jacob, "Sectorial Wages and the Real Exchange Rate," NBER Working Paper No. 1801, 1986.

Baldwin, Robert E., "The Political Economy of Protectionism," in Jagdish Bhagwati, ed., *Import Competition and Response*, Chicago: The University of Chicago Press, 1982.

______, (1984a) "Trade Policies in Developed Countries," in Ronald W. Jones and Peter B. Kenen, eds., *Handbook of International Economics*, Vol. I, Amsterdam: North-Holland.

______, (1984b) "Rent-Seeking and Trade Policy: An Industry Approach," *Weltwirtschaftliches Archiv*, *120*, 662–77.

Becker, Gary, "Investment in Human Capital," *Journal of Political Economy*, October 1962, *70*, 9–49.

Eaton, Jonathan and Grossman, Gene M., "Tariffs as Insurance: Optimal Commercial Policy When Domestic Markets Are Incomplete," *Canadian Journal of Economics*, May 1985, *18*, 258–72.

Hillman, Arye L., Katz, Eliakim and Rosenberg, Jacob, "Workers as Insurance: Anticipated Government Assistance and Factor Demand," *Oxford Economic Papers*, forthcoming, 1987.

Krueger, Anne O., "Trade Policies in Developing Countries," in Ronald W. Jones and Peter B. Kenen, eds., *Handbook of International Economics*, Vol. I, Amsterdam: North-Holland, 1984.

Kydland, Finn E. and Prescott, Edward C., "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

Maskin, Eric and Newbery, David, "Disadvantageous Oil Tariffs and Dynamic Consistency," Harvard Discussion Paper No. 1219, March 1986.

Mayer, Wolfgang, "Endogenous Tariff Formation," *American Economic Review*, December 1984, *74*, 970–85.

______, **and Riezman, Raymond,** "Endogenous Choice of Trade Policy Instruments," Working Paper Series No. 85-8, January 1985.

Mussa, Michael, "Imperfect Factor Mobility and the Distribution of Income," *Journal of International Economics*, February 1982, *12*, 125–41.

Roberts, Kevin, "The Theoretical Limits to Redistribution," *Review of Economic Studies*, April 1984, *51*, 177–95.

Rodrik, Dani, "Tariffs, Subsidies and Welfare with Endogenous Policy," *Journal of International Economics*, November 1986, *21*, 285–99.

Rogoff, Kenneth, "The Optimal Degree of Commitment to an Intermediate Target," *Quarterly Journal of Economics*, November 1985, 1169–89.

U.S. Congress, Public Law 93-618. Trade Act of 1975, January 3, 1974.

The Cyclical Behavior of Marginal Cost and Price

By MARK BILS*

This paper examines the cyclical behavior of price/marginal cost margins for U.S. manufacturing. Short-run marginal cost is markedly procyclical. In most industries, output price fails to respond to the cyclical movement in marginal cost; so price/marginal cost margins are markedly countercyclical. My results contradict business cycle theories that explain low production in a recession by a high real cost of producing; they support theories that explain low production in a recession by the inability of firms to sell their output.

First principles state that demand shocks are partially smoothed in the short run by upward price movements along a supply curve. The short-run supply curve is strictly upward sloping because some factors are fixed with remaining factors subject to diminishing returns. Therefore short-run marginal cost is increasing in output. This applies as well to aggregate demand shocks. A generally high level of demand should be associated with a general rise in the real price of outputs, where by real price I mean relative to input prices (a low real wage) as well as relative to surrounding periods (a high real interest rate). Such price movements should partially stabilize cyclical movements arising from demand shocks. This is one basis for classical economists' view of the macroeconomy as largely self-calibrating. The other is the belief that input prices respond reasonably quickly to variations in their demands (wage flexibility). In the *General Theory*, J. M. Keynes diverged from the flexible wage view, but remained faithful to the classical view that prices move procyclically relative to wages (countercyclical real wages).[1] A long empirical literature, however, has failed to find countercyclical real wages. (P. T. Geary and John Kennan, 1982, review much of the evidence.)

An obvious explanation for procyclical real wages is that the cycle largely reflects aggregate supply shocks. This is a centerpiece of many "new classical" models (for example, Finn Kydland and Edward Prescott, 1982). If productivity or input supplies are procyclical, then marginal cost will be countercyclical and real wages procyclical.

The purpose of this paper is to examine whether short-run marginal cost is procyclical and, if it is, whether output prices respond to the cyclical movements in marginal cost. Traditionally, capital has been the factor viewed as fixed in the short run. More recently, however, a number of studies (beginning with Walter Oi, 1962) recognize that labor may be costly to adjust. I examine short-run marginal cost allowing employment to be quasifixed. Using two-digit level manufacturing data for after 1956, I estimate that a short-run increase in production-worker employment of 10 percent was associated with an increase in marginal cost of about 2.4 percent. Most of the rise in marginal cost is due to overtime payments, incurred because employment is not perfectly flexible. The same 10 percent expansion was associated with a negative movement in output price (where output price is the

*University of Rochester, Rochester, NY 14627. This paper is revised work from my thesis at MIT. I thank Olivier Blanchard, Stanley Fischer, James Kahn, David Romer, Julio Rotemberg, Robert Solow and a referee for their helpful suggestions. I am also grateful for the comments of seminar participants at MIT and the University of Rochester.

[1]In light of empirical evidence, Keynes (1939) later acknowledged that there may be reasons why output prices do not respond to procyclical movements in marginal cost. This view had been expressed earlier by Pigou (1927).

industry-specific GNP deflator). Together these results show that price/marginal cost margins decrease by about 3.3 percent for a 10 percent expansion.[2]

My results do not contradict either the disequilibrium or equilibrium view of the business cycle; they do, however, strongly contradict leading versions of each. The results are incompatible with Keynesian theories based on a fixed short-run labor demand curve (for example, the *General Theory*). These theories assume that nominal wages are less flexible than prices. When prices increase (faster than expected) the real cost of labor falls, decreasing marginal cost and causing employment and output to expand, thereby tracing out the Keynesian aggregate supply curve. The evidence, however, is that marginal cost relative to price is high, not low, in a boom. The results are not inconsistent with disequilibrium models that allow price as well as wage rigidity (for example, Robert Barro and Herschel Grossman, 1971), because it is mostly cost, not price, movements which generate the cyclical movements in markups.

The finding is also incompatible with competitive equilibrium stories of the cycle, as they require a constant price/marginal cost ratio (equal to one). In particular, if the cycle is primarily the response of the economy to supply shocks then marginal cost should behave countercyclically. The evidence is clearly consistent with market-clearing models in which the elasticity of goods demand behaves procyclically. Michal Kalecki (1938) noted that the cyclical behavior of wages and prices for the United Kingdom might be explained by a procyclical elasticity of demand. A number of theoretical justifications for procyclical elasticity are possible (for example, Arthur Pigou, 1927; Julio Rotemberg and Garth Saloner, 1984; Mark Bils, 1985).

[2]Below I consider the possibility that effective output prices are more procyclical than the data show because of either a) countercyclical price discounting that is missed by the price series I use, or b) procyclical delivery lags. I provide evidence that these factors cannot be nearly large enough to explain the countercyclical markups I find.

I. Approach to Calculating Marginal Cost

A necessary condition for cost minimization is that the relative marginal products of inputs be set equal to their relative costs. This implies that, at the cost-minimizing choice of inputs, the *marginal* cost of increasing output can be calculated simply as the cost of increasing input i, where we are free to choose i, to produce the marginal increase in output. For my purposes it is convenient to think in terms of varying average hours of work for production workers, holding employment of production workers and all other inputs constant at their optimal levels. Marginal cost is

$$(1)\quad MC = (d\text{Costs}/dH) \times (dH/dY)|Y^*, H^*, N^*, \text{etc.}^*,$$

where Y is output, N is employment of production workers, H is average hours worked for production workers, and etc. are other inputs. The asterisk on Y implies it is chosen with regard to some overall objective (for example, profit maximization).

I assume throughout the paper that the productive technology has the form

$$(2)\qquad Y = H^{\alpha} f\,(\text{everything but } H),$$

which is slightly less restrictive than Cobb-Douglas. I note in particular that productivity shocks are allowed; however, they must be multiplicative with respect to average hours. This production function implies

$$(1')\quad MC = (1/\alpha)(H^*/Y^*)(d\text{Costs}/dH)|_*,$$

where the asterisk is shorthand for "at the optimum."

The standard macroeconomic approach is to set the marginal cost of an hour of labor equal to a wage rate W, and the cost of increasing average hours to employment, N, times that wage rate. (There are exceptions, including Andrew Abel, 1978; Matthew Shapiro, 1984; and Ben Bernanke, 1985.) Note, however, that this requires the marginal cost of an hour of labor to be invariant to the level of hours, H. Within manufacturing

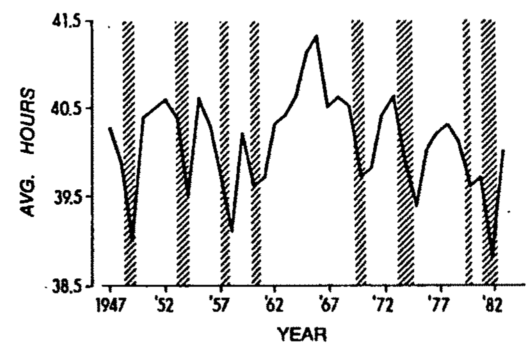


FIGURE 1. AVERAGE HOURS IN MANUFACTURING

this cost increases significantly with hours because firms are required to pay an overtime premium of 50 percent for hours above 40 per week (Fair Labor Standards Act of 1938). Even in industries not required to pay overtime premium, the marginal disutility of work presumably increases with the level of hours. If firms must compete for labor, compensation must reflect this higher disutility at higher hours.

This suggests viewing the effective cost of an hour of labor, W, as being a function of the number of hours worked, $W(H)$. Marginal cost of output then becomes[3]

$$(3)\quad MC=(1/\alpha)(H^*/Y^*)\times[W(H^*)N^*+W'(H^*)N^*H^*]=(1/\alpha)(N^*H^*/Y^*)\tilde{W}(H^*),$$

where $\tilde{W}(H^*)=W(H^*)+W'(H^*)H^*$.

$\tilde{W}(H)$ is interpreted as the "marginal wage schedule." For calculating marginal cost, it is clearly the marginal wage that is relevant. The arguments above predict $\tilde{W}(H)$ is increasing in H. If hours are procyclical, then this is potentially an important procyclical component in marginal cost. Average hours for production workers in manufacturing are given in Figure 1. The data source is the Bureau of Labor Statistics (BLS) *Employment and Earnings*. NBER-defined recessions (peak to trough) are shaded. Hours are markedly procyclical.

Procyclical movements in overtime hours will of course affect average wage rates over the cycle. This effect is small, however, because an average wage rate divides the overtime premium by all hours worked. For instance, below I estimate that, through an overtime premium, an increase in average hours per week from 40 to 41 raises the marginal wage by about 4.6 percent, but the average wage by only about 0.5 percent. Prior studies of the cyclical behavior of price markups have typically used average variable cost as a proxy for marginal cost (for example, Ian Domowitz et al., 1986). This explains why, compared to these studies, I find marginal cost to be much more procyclical and markups to be much more countercyclical.

If the marginal wage increases significantly with hours, then the question arises of why variations in hours would be observed. That is, why would firms not keep hours constant and vary labor input by varying employment? The answer must be that employment is less than perfectly flexible. If firms view large variations in employment as costly, it will be optimal for them to bear some costs of having hours away from their optimal long-run value. Quasi fixity of labor is studied in a number of papers (Oi, 1962; Ishag Nadiri and Sherwin Rosen, 1969; Frank Brechling, 1975; Thomas Sargent, 1978; Robert Pindyck and Rotemberg, 1983; Shapiro, 1984). For empirical purposes it is captured by including a convex function (typically quadratic) of change in employment in the firm's overall cost or profit function. The theoretical case for such adjustment costs is rarely made. For capital, convex adjustment costs have been justified on the grounds that increases in capital become increasingly difficult for the firm to absorb (Arthur Treadway, 1971) or, alternatively, that the cost of capital goods is an increasing function of the rate of investment (J. P. Gould, 1968). With regard to labor, the convex function of change in employment

[3] I ignore any potential adjustment costs for hours. A priori such costs seem unimportant for manufacturing because shifts and overtime hours appear easily adjustable. Moreover, studies that have examined adjustment costs for hours (Thomas Sargent, 1978; Shapiro, 1984) find them to be very insignificant.

can probably be better viewed as simply proxying for the fact that, for any of several good reasons, firms prefer a steady level of employment. Variations in employment require hires and fires (or recalls and layoffs), which firms may view as costly. The existence of firm-specific skills causes firms to prefer steady employment levels because decreases in employment cause trained workers to leave or have their skills depreciate, requiring additional training when employment is expanded. Because workers prefer stable employment, varying employment is also costly to a firm by hurting its reputation in the labor market.

Equation (3) gives marginal cost in terms of the cost of marginally increasing average hours holding other inputs, including employment, at their cost-minimizing values. Marginal cost can equivalently be viewed as the cost of marginally increasing employment holding other inputs, including average hours, at their cost-minimizing values. One component of this cost will be the marginal adjustment cost of increasing employment, given that employment is quasi fixed. Each of the rationales for quasi fixity of employment made above implies that this marginal adjustment cost will be high when employment is high relative to surrounding years. Thus the earlier statement that marginal cost will be high when hours are high can be reinterpreted as marginal cost is high when employment is relatively high.

II. The Marginal Wage

In the calculation of marginal cost in (3), α is a constant parameter and thus has no cyclical effect on marginal cost; and (NH/Y) are data.[4] Therefore, the problem of estimating cyclical movements in marginal cost reduces essentially to estimating the shape of the marginal wage schedule, $\tilde{W}(H)$. This is the major focus of the paper. Allowing for an overtime premium, the effective wage a firm faces is

$$W(H)=w[1+p(V/H)], \tag{4}$$

where w is the straight-time wage, p is the overtime premium, and V is the average number of overtime hours per production worker. The marginal wage with respect to hours is

$$W(H)=w[1+p(dV/dH)]. \tag{5}$$

For manufacturing industries (which constitute the present data), law dictates that an overtime premium be paid for hours above 40 per week. If all workers work an identical number of hours, then the relationship between average overtime hours and average hours would be very simple. V would equal H minus 40, or zero, whichever is larger. The change in overtime hours for a change in hours (dV/dH) would be equally simple, equaling zero when H is less than 40 and one when H is greater than 40. Provided the overtime premium is significant (such as the legal 50 percent rate), the implied marginal wage is very procyclical as hours typically move from below to above 40 during an expansion (note Figure 1).

However, the assumption that all workers work an identical number of hours is probably very poor. For instance, within a firm in which workers average 40 hours per week, one would expect to find some workers working more than 40 hours per week and some working less because of bottlenecks of one sort or another. Assuming that all workers work the same number of hours is particularly misleading when applied to data that are aggregated across time or firms, such as that used below. In a firm that averages 40 hours per week over a year (or quarter or month), part of the period is probably spent above 40 hours and part below; and in an industry averaging 40 hours per week, there are probably some firms above 40 hours and some below.

For this reason, my approach in calculating the marginal wage is to assume there is

[4] The data on hours H are for hired hours, whereas what is ideally needed are data on utilized hours. Given short-run fixity of some factors, it will in general be optimal to utilize hired labor more intensively in booms; so hired hours will be less procyclical than utilized hours. Therefore, by using data on hired hours rather than utilized hours, I cause productivity to appear more procyclical, and marginal costs less procyclical, than is true. Thus the bias works counter to my conclusions.

some variance in hours across workers. This implies that overtime hours are a smooth function of average hours. It remains true that the change in overtime hours for a change in hours (dV/dH), and therefore the marginal wage, should be increasing in H, and thus procyclical. When average hours are low, they can presumably be increased without much need for overtime hours; but when H is already high, a further increase in H will be difficult to achieve without a corresponding rise in overtime hours because virtually everyone will be working 40 or more hours per week. This implies a marginal wage that is continuously increasing in average hours rather than jumping at 40 hours per week.

Proceeding, a firm's overtime hours per worker can be exactly written as a function of the firm's average hours per worker and the higher moments of the distribution of hours across workers, Z.

$$V=f(H,Z). \tag{6}$$

How many moments must be included in Z depends on the nature of the distribution of hours. V should be increasing in H and also increasing in the variance of hours across workers. The firm could clearly reduce its overtime costs by reducing the variance of hours across workers to zero. Presumably firms arrive at an optimal variance by trading off overtime costs against the inconvenience (or bottlenecks if workers have differing assignments) of scheduling hours so that all work the same number. Fortunately, I need not consider this portion of the firm's problem explicitly, because I can calculate marginal cost in terms of increasing average hours given the firm's choice for the variance in hours across workers.

Of primary interest is the derivative of overtime hours with respect to average hours. Following the arguments above, I specify this derivative to be increasing in average hours. I assume

$$dV/dH=a+b(H-40)+c(H-40)^2 + d(H-40)^3+x. \tag{7}$$

The parameter a gives the increase in overtime hours for an increase in hours at hours equal to 40 per week. If hours are symmetric around H, we should expect a to equal about 0.5. The key parameter of interest is b. It describes how rapidly the derivative of overtime hours with respect to hours rises at H equal to 40 per week. In turn, it will determine how steep the marginal wage schedule is at H equal to 40 per week. In estimating, I consider the possibility that both parameters a and b vary across industries and vary with time trends. As H approaches zero no one will be working 40 hours per week and (dV/dH) should approach zero. On the other hand, as H becomes very large everyone will be working overtime and (dV/dH) should approach one. I try to capture this nonlinear relationship with the terms $(H-40)^2$ and $(H-40)^3$. The coefficient on the cubic term, d, is expected to be negative. The quadratic term has no predictable sign. x is a potential error term in the relationship. I try three estimators of equation (7) below. The first requires x exactly equal to zero; the second allows x to be nonzero, but requires it be uncorrelated with H; the third allows x to be nonzero and potentially correlated with H. This issue is discussed further below.

Given (7), the marginal wage schedule is

$$\tilde{W}(H)=w\left[1+pa+pb(H-40) + pc(H-40)^2+pd(H-40)^3+px\right]. \tag{5'}$$

This marginal wage specification is similar to those of Abel (1978) and Shapiro (1984), although they exclude the higher-order terms $(H-40)^2$ and $(H-40)^3$.

To this point I have treated the overtime premium, p, as a parameter to be determined. This may seem curious given that law prescribes a premium equal to 50 percent. It may be incorrect, however, to equate the *effective* premium for overtime hours with the 50 percent rate firms must pay by law.[5] Several papers, most notably Robert

[5]Beyond the discussion in the text, the effective premium would be less than 50 percent if firms circumvented the legally required premium. This does not

Hall (1980a), have pointed out that if a long-term relationship exists between employer and employees, then it is not correct to equate the effective wage rate with the wage payment being made at any single point in time. If a firm pays its workers in excess of the marginal disutility of labor in one period, it may well be tied with paying them less than the marginal disutility of labor in another period. With regard to overtime payments, the problem is the following. The wage a firm pays takes a very large jump of 50 percent at 40 hours per week due to the overtime premium. Workers' disutility of working is presumably smoothly increasing in hours. This implies workers would strictly prefer working some overtime hours to working 40 hours per week (in fact, overtime hours are rationed in many instances). By offering workers overtime hours, therefore, a firm may incur some goodwill, which allows it to lower compensation in another form, if not then, at some other time. The implication is that the effective cost premium of an overtime hour may be less than the 50 percent explicit payment. At a sufficiently high level of overtime hours, workers will presumably disapprove of any increase in hours, despite the overtime pay. At this point the effective premium would exceed 50 percent.

In Section III, two separate approaches are used for estimating the marginal wage. The first estimates the effect of hours on overtime hours directly. To arrive at a marginal wage, I then must presume the effective overtime premium equals 50 percent. The second approach estimates the shape of the marginal wage schedule indirectly from observing the cost-minimizing choices firms make for employment and average hours. This approach simultaneously estimates the effect of hours on overtime hours and the overtime premium. Therefore, it is not necessary to presume that the effective overtime premium equals 50 percent; and the second approach is much less subject to the criticism just raised. I find the two approaches give fairly similar estimates for the shape of the marginal wage. This supports an assumption that the effective overtime premium is near 50 percent.

Even estimating the marginal wage by the second approach requires assuming that variations in hours affect compensation through an overtime premium. If the model of labor payments as "installment payments" is taken to its logical end, then this assumption will be false. The effective cost of an hour of labor will be the marginal disutility of an hour of labor; the distinction between straight-time and overtime hours becomes irrelevant, and my estimates of the marginal wage below fail to be identified. This is not as dire a problem as it sounds, however. If the effective marginal wage equals the marginal disutility of labor, then its shape can be inferred from prior studies' estimates of labor supply elasticities. To obtain a marginal wage that is less procyclical than I find below requires that the elasticity of labor supply be greater than 0.72.[6] This is above most estimates in the literature. (Hall, 1980b, reviews results from several studies.) Therefore, redoing the present paper equating the shape of the marginal wage to the shape of workers' work/leisure indifference curves would imply an even more procyclical marginal cost and more countercyclical markup than I purport to show.

III. Estimating the Marginal Wage

As mentioned, I try two separate approaches to estimating the marginal wage. The first estimates the relationship between overtime hours and hours. Given a value for

appear to be an issue. A 1965 survey of workers found that among laborers in manufacturing 83.3 percent of those working overtime received premium pay (U.S. Department of Labor). More important, the data on overtime hours I use below, which is gathered by the BLS, define overtime hours as those weekly hours that exceed regular hours and for which overtime premium is paid.

[6]From (5′) the elasticity of the marginal wage with respect to hours H is approximately $[pb+2pc(H-40)+3pd(H-40)^2]\ H/(1+pa)$. Using the same estimates for pa, pb, pc, and pd as are used in Section IV, the average of this elasticity for the sample is 1.39. The inverse of this elasticity, 0.72, is the number stated in the text as corresponding to an estimate of elasticity of labor supply.

the overtime premium of .5, this implies the marginal wage. The second estimates the shape of the marginal wage schedule from observing the cost-minimizing choices firms make for hours and employment. I find the two approaches yield fairly similar estimates for the shape of the marginal wage schedule.

A. *Approach* 1

The goal here is to estimate (dV/dH) directly. In turn, this will identify the marginal wage schedule, assuming an overtime premium equal to .5. Totally differentiating equation (6) gives

$$(8)\quad dV=(dV/dH)\,dH+(dV/dZ)\,dZ,$$

where all variables are as previously defined; and it is understood that all variables refer to a given time period t. Substituting from (7),

$$(8')\quad dV=\left[a+b(H-40)+c(H-40)^2+d(H-40)^3+x\right]dH+(dV/dZ)\,dZ.$$

I intend to estimate this equation in order to identify the parameters a, b, c, and d. The error term x is problematic here because it yields a nonconstant parameter for (dH) (beyond that captured by the parameters a and b varying over industries and time). Therefore, for this subsection only, I impose that x be equal to zero. The higher moments of the distribution of hours Z are unobservable. This creates problems in estimating (8′) only to the extent that (dZ) is correlated with (dH). I proxy for (dZ) with a constant, time trends, the rate of growth in employment, and the rate of acceleration in employment. This obviously does not capture all of (dZ); but I assume the remainder of (dZ) is uncorrelated with (dH). Substituting in (8′) yields

$$(8'')\quad dV=\left[a(i,t)+b(H-40)+c(H-40)^2+d(H-40)^3\right]dH+g(t)+k\ln(N/N_{-1})+m\left[\ln(N/N_{-1})-\ln(N_{-1}/N_{-2})\right]+e.$$

TABLE 1—ESTIMATES OF EQUATION (8″)[a]

$a=$	.3855 + .0314 $(t-1956)$ − .00106 $(t-1956)^2$
	(10.98) (5.41) (−5.42)
$b=$	.1208
	(6.08)
$c=$	−.00431
	(−1.04)
$d=$	−.003203
	(−2.00)
$g=$	−.0824 + .0395 $(t-1956)$ − .00321 $(t-1956)^2$ + .000070 $(t-1956)^3$
	(−2.43) (3.88) (−3.87) (3.58)
$k(i)=$	1.100
	(4.79)
$m(i)=$	.640
	(2.84)
$R^2=$	.925
$D\text{-}W=$	2.44
$F=$	85.6
$n=$	563

Note: t-statistics in parentheses.

[a] F-tests dictate allowing the parameters k, m, and the linear time trend in a to vary across industries. The table reports mean parameters. F-tests for whether the intercept of a and b vary by industry respectively equal 1.25 and 1.17; the value for rejecting a constant parameter with 95 percent confidence is 1.6.

The error term, e, reflects the uncaptured effect of (dZ).

I estimate (8″) using BLS *Employment and Earnings* data on annual averages for overtime hours, hours, and production-worker employment for each of the 21 SIC classified two-digit manufacturing industries for each year from 1956 to 1983. (BLS collection of overtime data began in 1956.) The 21 industries are listed in Table 5. Transportation equipment is broken into motor vehicles and equipment, and the remainder of transportation equipment. For these two industries the coverage begins with 1958. Several variables require first differencing the data; so the coverage is actually for 1957 to 1983. For hours H, I use the average of hours in years t and $t-1$.

The results of estimating (8″) by OLS appear in Table 1. The estimate for the parameter a (using the mean across industries as of 1956) implies that an increase in average hours from 40 to 41 hours per week is associated with an increase in aver-

age overtime hours of .386 hours. The parameter a depends significantly on linear and quadratic time trends. After 1956 an increase in average hours is associated with a greater increase in overtime hours. (The peak effect is reached in 1971; an increase in H from 40 to 41 raises average overtime hours by about .618 hours.) I considered allowing the parameter a to vary by industry; but the data support restricting it to be equal across industries. (See bottom of Table 1.)

The key parameter is b. Parameter b and, to a lesser extent, parameters c and d dictate the shape of the marginal wage schedule. The estimates for b, c, and d imply that, whereas an increase in average hours from 40 to 41 increases overtime hours by .386 hours in 1956, an increase from 41 to 42 hours increases overtime hours by .499 hours. Given an overtime premium of 50 percent, this translates into about a 4.6 percent higher marginal wage at 41 hours than at 40. (Recall Figure 1 showed post-World War II recessions to be associated with a fall in average hours of about 1 hour.) By contrast, the increase in overtime hours between 40 and 41 hours per week would only raise an *average* wage rate by approximately 0.5 percent (or one-ninth the rise in the marginal wage). I also considered allowing the parameter b to vary across industries and with time trends; but the data support restricting it to be a constant. (See bottom of Table 1.)

Figure 2 presents graphically the relationship between hours and overtime hours implied by the estimates for the parameters a, b, c, and d. Over the range of 36 to 43 average hours per week, which is approximately the range observed in the two-digit industry data,[7] (dV/dH) increases from .039 to .623 for 1956, from .152 to .736 for 1971, and from .113 to .670 for 1983.

This first approach to estimating the marginal wage has the virtue of being very simple and straightforward. It has two drawbacks. It does not allow an error in the relationship between hours and overtime hours–equation (7). More important, it requires that the effective overtime premium equal the explicit premium of 50 percent.

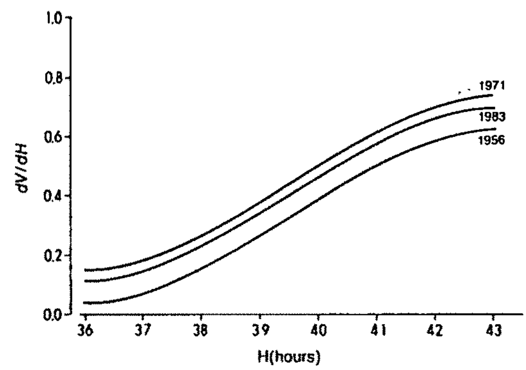


FIGURE 2. RELATIONSHIP BETWEEN H AND dV/dH

B. *Approach* 2

The shape of the marginal wage schedule can be inferred from the choices firms make for hours and employment. For estimating the absolute value of marginal cost, it is necessary to know the absolute value of the marginal wage schedule; but for estimating relative cyclical movements in marginal cost, all that is needed is the shape of the marginal wage schedule.

I consider the firm's problem of choosing employment and average hours to minimize the cost of its desired quantity of production-labor, L^*. For a firm to minimize its overall cost function or to maximize profits, it must minimize this cost. Therefore, the problem is more general than profit maximization or cost minimization.[8]

[7]96 percent of the data observations for hours fall between 36 and 43 hours per week.

[8]Looking at the firm's broader cost-minimization problem would provide further first-order conditions for capital, nonproduction labor, materials, and other inputs. Although estimating a set of conditions theoretically should yield more efficient estimates, it is unlikely that these further first-order conditions can be consistently estimated. Estimating conditions for capital and nonproduction labor is difficult because accurate measures of short-run movements in *utilized* capital and nonproduction labor are not available. Because capital and nonproduction labor are largely fixed in the short run, it will generally be optimal for firms to utilize these factors more intensively in a boom. Therefore, data on hired capital and nonproduction labor will have a con-

Because I wish to allow for the possibility that employment is quasi fixed, I actually consider the firm's dynamic problem of minimizing the expected present-discounted value of the costs of procuring its expected future stream of production-labor demands. This problem is

$$(9)\quad \underset{N_t, H_t}{\mathrm{Min}}\, E_t \sum_{\tau=t}^{\infty} R_{t,\tau}\{W_\tau(H_\tau)N_\tau H_\tau + F_\tau N_\tau + qw_\tau N_\tau H_\tau[\ln(N_\tau/N_{\tau-1})]^2\}$$

subject to

$$E_t\{N_\tau H_\tau^\beta - L_\tau^*\} = 0, \quad \text{for } \tau = t, t+1,\ldots,.$$

E_t is the expectations operator, conditioned on information known at time t. I choose an infinite horizon for simplicity. $R_{t,\tau}$ is the nominal rate of discount between periods t and τ.

$W(H)NH$ is the wage cost of production labor. The variable F captures production-labor expenses that increase with employment but are fixed with respect to average hours, such as unemployment insurance and vacation pay. The real-world distinction between hours-related and not-hours-related expenses is discussed at length below. Equation (9) assumes that firms possess no monopsony power with respect to employment; however, the effective wage firms must pay is affected by their choice of hours.

Fixity of employment is introduced, similarly to other works, by incorporating a convex function of change in employment. The costs of adjustment are of the external variety. (A discussion of various forms for adjustment costs is contained in Hans Soderstrom, 1976.) Adjustment costs are often specified as quadratic in changes (that is, $(N_t - N_{t-1})^2$). The formulation here makes the marginal adjustment cost linear in percentage changes rather than absolute changes in employment. This is desirable because the two-digit industries vary considerably in size; and it does not seem reasonable that an increase in employment of 1 million in a large industry such as primary metals would have the same relative effect on costs as a 1 million increase in a small industry such as leather goods. The adjustment costs are multiplied by the straight-time wage to account for nominal movements (as in Pindyck and Rotemberg).[9]

The constraint in (9) requires that hours and employment be sufficient to satisfy the labor input demand, L^*. The effective amount of labor is given by NH^β. This specification does not require that employment and hours enter multiplicatively in production. I have not set β equal to one because several studies have found that increasing hired labor by increasing hours has a greater impact on output than an equal increase in hired labor achieved by increasing employment.[10]

To meet (11), firms must satisfy a dynamic first-order condition that the marginal cost of an hour of effective labor (NH^β) obtained by increasing average hours H, equal the marginal cost of an hour of effective labor obtained by increasing employment N, including a marginal adjustment cost. This first-order condition can be written:

$$(10)\quad E\{W(H)+(F/H)+q[w\ln(N/N_{-1}) - Rw_{+1}(N_{+1}/N)\ln(N_{+1}/N)]\} = (1/\beta)\tilde{W}(H).$$

Time period t subscripts have been dropped for convenience. By estimating this condition I obtain an estimate of the marginal wage up to the multiplicative constant $(1/\beta)$.

siderably cyclical bias. A further problem with estimating first-order conditions for capital, nonproduction labor, or materials is that it requires that these factors can be substituted for one another and for other factors. This seems questionable for the short run. The assumption of short-run substitution between hours and employment of production workers, on the other hand, is not stringent.

[9] I ignore any adjustment costs for hours. See footnote 3.

[10] For instance, Shapiro (p. 47) finds hours have about a 6 percent higher marginal product than employment.

For reasonably small changes in employment, such as those observed in the annual, two-digit industry data used below, $[(N_{+1}/N)\ln(N_{+1}/N)]$ can be closely approximated by $[\ln(N_{+1}/N)]$. Making this simplification, substituting for the marginal wage from (5′), and rearranging gives

$$(10') \quad E\{\ln(N/N_{-1}) - R(w_{+1}/w)\ln(N_{+1}/N)\} = (-1/q)[(W(H)/w)+(F/wH)] + (1/q)(1/\beta)\big[1+pa+pb(H-40) + pc(H-40)^2+pd(H-40)^3+px\big].$$

Examining (10′) it is clear that none of the parameters β, p, a, b, c, or d are identified. The values of pb, pc, and pd relative to $(1+pa)$ are identified, however, and this gives the shape of the marginal wage schedule.

Equation (10′) as stands cannot be estimated because it includes an expectational term. Substituting the actual value for the expected gives

$$(10'') \quad \ln(N/N_{-1}) - R(w_{+1}/w)\ln(N_{+1}/N) = (-1/q)[(W(H)/w)+(F/wH)] + (1/q)(1/\beta)\big[1+pa+pb(H-40) + pc(H-40)^2+pd(H-40)^3\big]+\eta,$$

where

$$\eta = (p/q\beta)x + (Rw_{+1}/w)\ln(N_{+1}) - E\{(Rw_{+1}/w)\ln(N_{+1})\}.$$

η is a composite error term. The first component reflects a potential error in the relationship between hours and overtime hours, (dV/dH). The second component is an expectational error. If expectations are rational, it will be uncorrelated with all variables of which firms have knowledge at time t, including the right-hand side variables. Thus, if there is no error in the (dV/dH) relationship (x equals zero), or if the error is orthogonal to the right-hand side variables in equation (10″), then equation (10″) can be consistently estimated by ordinary least squares. More generally, we should expect the error term x to be correlated with the right-hand side variables.[11] Therefore, I estimate (10″) both by OLS and by instrumental variables (described more fully below).

I estimate (10″) using annual, post-World War II, two-digit manufacturing data. The source for employment N, average hours H, average hourly wage $W(H)$, and average straight-time wage w is the BLS *Employment and Earnings*. Availability of data on the straight-time wage restricts this data to 1956 and after. For the nominal discount rate, I use the prime rate charged by banks; the source is the Board of Governors of the Federal Reserve.

The variable (F/wH) requires some discussion. Since 1951 the U.S. Chamber of Commerce has surveyed firms on their annual fringe payments and legally required payments to workers. The surveys were conducted biannually through 1977, and annually since. There is not a direct relationship between all such payments and F because some payments, such as FICA payments, vary with hours as well as employment. By deleting such payments, a measure of F is obtained.[12] The Chamber *Survey*

[11] A possible source of disturbance to the (dV/dH) relationship would be shocks to the *elasticity* of labor supply with respect to hours. Standard shocks to labor supply (in which labor supply shifts out proportionately so that workers are willing to accept a lower real wage at any given level of hours) are already captured by the straight-time wage. An increase in elasticity would correspond to workers being willing to work longer hours without a corresponding rise in overtime premium. This would imply a negative value for the error term x and would presumably lead firms to expand H. (Such shifts in elasticity are, however, irrelevant if firms treat the legal premium as the effective premium.)

[12] The Chamber of Commerce survey has information on 22 types of payments, which are then subsumed into 5 major categories. The first major category is legally required payments. These payments increase with earnings for a given worker up to a ceiling, at which point they become fixed with respect to hours. FICA

gives fringe payments as a percentage of wages for all workers. By using their figures, I implicitly assume that (F/wH) for production workers can be represented by (F/wH) measured for all workers. The Chamber *Survey* gives fringe payments as a percentage of wages. (F/wH) is fixed payments in terms of straight-time wages. Therefore, I put the Chamber's figures in terms of straight-time wages by multiplying by $(W(H)/w)$. Although the Chamber's *Survey* covers all manufacturing industries, it does not give separate figures for each industry. Lumber is combined with furniture and with paper, instruments with miscellaneous manufactures, textiles with apparel, foods with tobaccos, and rubber and plastics with leather goods. Nevertheless, there are 14 categories with 16 observations for each (1951 to 1981 biannually, plus 1978, 1980, and 1982), for a total of 224 observations. The other variables in equation (10″) were aggregated into the same 14 categories in order to be consistent with the variable (F/wH).

The results of estimating equation (10″) by OLS appear in Table 2a. The parameter estimates are in accord with the framework above and are statistically significant. The estimate for the adjustment-cost parameter q implies that when employment is 10 percent above surrounding years, the marginal adjustment cost of an additional worker is approximately 495 dollars in 1967 (using the mean wage and hours for manufacturing). In 1967 annual compensation in manufacturing averaged about \$7,480; so the marginal adjustment cost is about 6.6 percent of annual compensation. In a further regression (not shown), I allowed the parameter q to vary across industries by making it linearly related to a two-digit industry measure of labor turnover;[13] but this had a very insignificant effect on the results.

Of central interest is the estimated shape of the marginal wage schedule. The relative steepness of the schedule is given by the value of pb, pc, and pd divided by $(1+pa)$. I considered allowing both parameters a and b to vary with time trends and across industries. F-tests dictated allowing the parameter a to vary, but restricting the parameter b to be a constant. The estimates for pb, pc, and pd relative to $(1+pa)$ imply that, for the year 1956, the marginal wage at average hours equal to 41 is about 6.6 percent higher than the marginal wage at average hours equal to 40. (In this calculation I am using the mean industry value for a. For later years, the estimated marginal wage schedule is less steep because pa is larger.) This is a little steeper than the marginal wage schedule estimated with the first approach assuming an overtime premium of 0.5. There, going from H equal to 40 to H equal to 41 raised the marginal wage by about 4.6 percent. The estimates obtained by the two approaches, however, would not be significantly different statistically.

I reestimated equation (10″) by instrumental variables. The instruments are $(H-40)$, $(H-40)^2$, $(H-40)^3$, and $[(W(H)/w)+(F/wH)]$, each lagged two years, the rate of growth in domestic consumer credit in

taxes and workers' compensation payments have ceilings that are considerably higher than average earnings. Therefore, I consider these payments to increase with hours and do not include them in F. (For instance, for 1979 the ceiling for FICA payments equaled 178 percent of average earnings, and workers' compensation had effectively no ceiling. The source here is Hart, 1984, p. 15.) Unemployment compensation has a ceiling considerably lower than average earnings. Therefore, I consider these payments to be unrelated to hours and include them in F. (For 1979 the ceiling equaled 47 percent of average earnings.) The final type payment in the first category, Railroad Retirement taxes, is of no significance. I treat it as hours related. The second category includes private pension plans, and insurance and other benefits. I consider these to be unrelated to hours and include them in F. The third and fourth categories are payments for time not worked, primarily paid lunch or break time, paid sick leave, and paid vacation time. I consider these to be unrelated to hours and include them in F. The final major category is other items. These are primarily profit-sharing payments and special bonuses. I consider these payments to be either hours related or not relevant to production workers. In neither case do I include them in F.

[13] I measured industry turnover by the average of industry separations per 100 employees for the years 1967 to 1969. These three years were chosen because they are at the middle of the sample period and because they are relatively noncyclical. The data source is the BLS' *Employment and Earnings*.

TABLE 2—ESTIMATES OF FIRST-ORDER CONDITION

	OLS Estimates (1)	IV Estimates (2)
$q=$	.8618	.5381
	(4.63)	(2.86)
$(1/\beta)[1+pa(t)]=$	$1.2881-$ $.0342\,(t-1956)$	$1.2487-$ $.0235\,(t-1956)$
	(337.15) (−4.05)	(440.28) (−2.79)
	$+.00281\,(t-1956)^2$	$+\ .00206\,(t-1956)^2$
	(4.63)	(3.41)
	$-\ .000054\,(t-1956)^3$	$-\ .000038\,(t-1956)^3$
	(−4.26)	(−2.84)
$(1/\beta)pb=$	.0938	.1161
	(4.40)	(2.27)
$(1/\beta)pc=$	−.0061	−.0101
	(−2.57)	(−1.38)
$(1/\beta)pd=$	−.00238	−.00500
	(−2.78)	(−1.33)
$\frac{pb}{1+a(1956)}=$	.0728	.0930
	(4.63)	(2.24)
$\frac{pc}{1+a(1956)}=$	−.0047	−.0081
	(−2.62)	(−1.37)
$\frac{pd}{1+a(1956)}=$	−.00185	−.00401
	(−2.83)	(−1.31)
$R^2=$	.43	.29
$\rho^2=$	.060	.178
	(0.81)	(2.57)
$F=$	7.23	3.43
$n=$	224	196

Note: t-statistics in parentheses.

years t and $t-1$ (source is Board of Governors of the Federal Reserve System), and the rate of increase in energy prices for total manufacturing in years t and $t-1$ (source is Ernst Berndt and David Wood, 1984). The credit and energy price variables are intended to reflect aggregate demand and aggregate supply shocks, respectively; each is deflated by the GNP deflator price index. The results are presented in Table 2b. The estimated adjustment cost parameter is reduced by about 37 percent. The estimated marginal adjustment cost of an additional worker when employment is 10 percent above surrounding years is now only 310 dollars in 1967, or about 4.1 percent of annual compensation. The estimated marginal wage schedule is a little steeper. An increase in average hours from 40 to 41 is associated with about an 8.1 percent higher marginal wage.[14]

The alternative estimators for the marginal wage schedule, although disagreeing on an exact slope, each suggest a schedule that responds significantly to average hours. In the next section I find that cyclical movements along this marginal wage schedule are

[14] I conducted a specification test of the OLS estimates of the marginal wage schedule by reestimating equation (10″) by instrumental variables constraining the parameters $pb/(1+pa)$, $pc/(1+pa)$, and $pd/(1+pa)$ to be equal to their OLS estimates. An F-test of this restriction equals 4.02; the critical value for rejection of the restriction with 99 percent confidence is about 3.9. So the restriction is statistically rejected, although the difference in the estimated schedules appears economically small.

large relative to cyclical movements in straight-time wage rates or output prices.

IV. The Markup

Given an estimate of the marginal wage, calculating marginal cost is straightforward. Substituting in (3) for $W(H)$ from (5′) gives

$$(11)\quad MC = w\big[1 + pa + pb(H-40) + pc(H-40)^2 + pd(H-40)^3 + px\big] \times (NH/Y)(1/\alpha),$$

where asterisks have been dropped for convenience. Taking logs and approximating for H near 40 gives

$$(12)\quad \ln(MC) = \ln(w) + \frac{pb(H-40) + pc(H-40)^2 + pd(H-40)^3}{1 + pa + px} + \ln(NH/Y) + (\text{intercept and trend terms}).$$

As expressed in equation (12), marginal cost has three components of interest: the straight-time wage, movements along the marginal wage schedule, and a productivity term. The error effect px is unobservable because it is combined in the estimates above with the expectational error. The similarity of the OLS and instrumental variable results, however, suggests that this term is either small in variance or cyclically insensitive. Therefore I ignore it in calculating marginal cost. Three potential estimates for the shape of the marginal wage schedule were presented in Section III. For substituting in (12) I choose the estimates from the first approach, which estimated the relationship between hours and overtime hours directly. The alternative estimates give somewhat steeper marginal wage schedules, and so would give results even more favorable to my conclusions.

Of principal interest is the price/marginal cost markup. Given (12), this is estimated by

$$(13)\quad \ln(P/MC) = \ln(P) - \ln(w) - \frac{p\hat{b}(H-40) + p\hat{c}(H-40)^2 + p\hat{d}(H-40)^3}{1 + pa} - \ln(NH/Y) - (\text{intercept and trend terms}),$$

where the carets signify the OLS estimates from equation (8″). I examine the cyclical behavior of markups by regressing the measure (13), component by component, on a measure of the business cycle. Because I use industry-level data, I need an industry-level measure of the cycle (as opposed to aggregate measures such as the NBER reference cycles). The measure I use is production-worker employment relative to the four surrounding years:

$$\ln(N) - 0.25\ln(N_{-2}N_{-1}N_{+1}N_{+2}).$$

The data are annual averages for each of the 21 two-digit manufacturing industries listed in Table 5. The sample period is 1956 through 1981. (This incorporates data for 1954 through 1983 on N, as the cyclical measure absorbs two leads and two lags.) As before, N, H, and w are from the BLS *Employment and Earnings*. Y is the industry-specific GNP from the U.S. Commerce Department, and is thus a value-added measure; P is the industry-specific GNP deflator.[15]

The results of regressing each component of the markup from equation (13) on the measure of the business cycle appear in Table 3. The regressions also include trends (t, t^2, t^3) and industry-specific constants. The regressions employ a Cochrane-Orcutt AR(1) correction. Looking at Table 3, beginning with row 2, straight-time wages show a very small countercyclical movement. A 10 per-

[15] I thank Michael Burda for making these data available.

TABLE 3—REGRESSIONS OF MARGINAL COST AND PRICE ON BUSINESS CYCLE

Component	Estimate
$\ln(P)$	$-.0873$ (-2.54)
$\ln(w)$	$-.0374$ (-2.92)
$\dfrac{pb(H-40)+pc(H-40)^2+pd(H-40)^3}{1+pa}$	.2078 (14.76)
$\ln(NH/Y)$	.0728 (2.18)
$\ln(P/MC)$	$-.3335$ (-9.17)

Note: T-statistics in parentheses; $n = 563$.

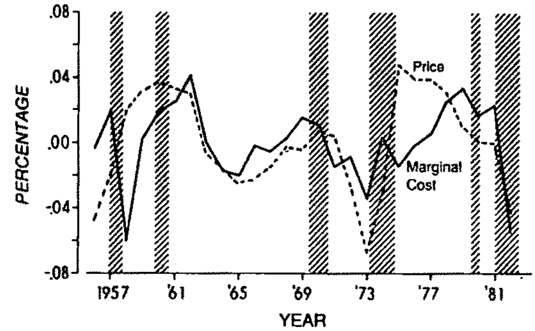


FIGURE 3. THE BEHAVIOR OF AGGREGATE MARGINAL COST AND AGGREGATE PRICE

cent short-run increase in employment is associated with a 0.4 percent decrease. Most of the action, however, is *within* the wage schedule $W(H)$. Looking at row 3, a 10 percent increase in employment increases the marginal wage 2.1 percent by moving up the marginal wage schedule. This component of the markup shows the most dramatic cyclical movements.

Productivity is slightly countercyclical. A 10 percent short-run increase in employment is associated with a 0.7 percent increase in (NH/Y). My finding of countercyclical productivity may seem surprising given that procyclical labor productivity is a noted empirical feature of cycles (see Victor Zarnowitz, 1985). In calculating marginal product, however, I have defined productivity differently from most studies of labor productivity. My measure of labor is more procyclical because it includes variations in average hours. Furthermore, I examine only production-worker labor; production labor is much more cyclical than nonproduction labor. By disaggregating I reduce the apparent procyclicality of productivity. Industries that decline most in recessions are those with higher labor productivity. Therefore, aggregate productivity is more procyclical than disaggregate.

Nominal marginal cost (combining rows 2, 3, and 4) is procyclical, increasing by about 2.4 percent with the 10 percent short-run increase in employment. Much of this procyclical movement is from the impact of average hours on the marginal wage.

Table 3 also gives the behavior of prices (GNP deflators). Prices are somewhat countercyclical. Prices decrease by about 0.9 percent for a 10 percent short-run increase in employment. Of primary interest, price/marginal cost margins are very countercyclical. Margins decrease by about 3.3 percent for the 10 percent increase.

Figure 3 graphs price and marginal cost for aggregate manufacturing. Aggregate marginal cost is constructed by aggregating the marginal costs for the individual industries, giving each industry a weight equal to its share in manufacturing value added in 1967. Aggregate price is constructed by aggregating industry-specific deflators in an identical fashion. Constant and trends (t, t^2, t^3) have been eliminated. The six NBER-defined recessions (peak to trough) during the sample period are shaded. The most dramatic movements in markups occur for the 1958 and 1974–75 recessions; in each case markups increase by about 10 percent. The recessions of 1970–71 and 1982 are each associated with markup increases of between 3 and 4 percent. The two relatively mild recessions of 1961 and 1980 show little or no effect on markups. The average increase in markups for the six recessions is about 4.6 percent.

The price data (GNP deflators) I use are constructed from sellers' reported prices. If price discounting is more prevalent in recessions, then this will be a biased measure of

TABLE 4—STIGLER-KINDAHL RESULTS

	BLS	NBER
Trend		
Monthly Percentage Rate of Increase	−.026	−.060
Cycle		
Average Monthly Percentage Rates of Change		
Peak to Trough	−.129	−.205
Trough to Peak	.118	.079
Average Monthly Percentage Rates of Change Corrected for Trend		
Peak to Trough	−.082	−.140
Trough to Peak	.117	.111
Short Run		
Correlation of First Differences of Logarithms		
Monthly	.378	
Quarterly	.576	
Semiannually	.728	
Variances of First Differences of Logarithms	.202	.042

Source: Stigler and Kindahl, 1970, p. 82.
Note: NBER prices are Stigler and Kindahl prices. Comparison of the Comprehensive BLS Index with the Corresponding NBER Index for all Industrial Commodities.

cyclical price variability, with actual transaction prices being more procyclical than my data appear. The most extensive study of this issue, to my knowledge, is by George Stigler and James Kindahl (1970). Stigler and Kindahl collected purchase-price data for a large number of industrial goods for the years 1957 to 1966. They find that movements in their transaction prices differ considerably from movements in BLS Wholesale prices (an index constructed from sellers' listed prices) at a monthly frequency. At a cyclical frequency, however, their prices and the BLS prices behave roughly similarly. Results from their study are reprinted in Table 4. The Stigler-Kindahl transaction prices decline more than BLS prices in recessions, but actually increase somewhat less in expansions. Although the correlation between Stigler-Kindahl price changes and BLS price changes is only 0.38 at a monthly frequency; at a 6-month frequency the correlation is 0.73. At an annual frequency (the data frequency here) the correlation is presumably even higher. My conclusion is that the failure of prices to move with marginal cost probably cannot be explained by failure of reported prices to move with transacted prices.

Effective prices charged may be more procyclical than the data show if delivery lags are procyclical. (Dennis Carlton, 1979, has considered markets that use both price and delivery lag movements to clear.) In a wholesale market a delivery lag raises the effective price to the buyer by delaying the markup received by the buyer on that good. Therefore, the impact of cyclical movements in delivery lags depends on the size of markups on wholesale goods and on the rate-of-time discount. I examined the behavior of delivery lags for total manufacturing for the years 1956 to 1982. (Delivery lags, in months, are approximated by the ratio of unfilled orders to monthly deliveries. The data source is the Commerce Department.) Looking at peaks and troughs of the six NBER-defined recessions, I find that delivery lags averaged 3.56 months at the peaks and 3.38 months at the troughs—a difference of 0.18 months. For plausible markups and real interest rates these movements in delivery lags cannot possibly explain the cyclical movements I find in markups. For instance, suppose a markup equal to the price of the wholesale good and a real interest rate of 10 percent; then a delivery lag increase of 0.18 months would raise the effective price by about 0.15 per-

TABLE 5—REGRESSIONS OF MARGINAL COST AND PRICE ON CYCLE, BY INDUSTRY

	$\ln(P)$	$\ln(w)$	Hours Effect[a]	$\ln(NH/Y)$	$\ln(P/MC)$
Lumber and Wood Products	.475 (2.3)	−.050 (−1.1)	.232 (4.1)	.307 (1.6)	.005 (0.1)
Furniture and Fixtures	−.265 (−3.8)	−.069 (−2.0)	.319 (4.7)	−.062 (−0.6)	−.447 (−5.7)
Stone, Clay and Glass Products	−.161 (−1.1)	−.055 (−1.2)	.279 (5.0)	.102 (1.0)	−.481 (−4.2)
Primary Metals	−.313 (−2.3)	−.105 (−1.7)	.323 (2.9)	−.226 (−2.6)	−.209 (−1.2)
Fabricated Metals	−.330 (−3.3)	−.115 (−2.4)	.239 (4.3)	.047 (0.6)	−.486 (−6.7)
Machinery except Electrical	−.185 (−2.3)	−.091 (−2.6)	.239 (4.2)	.187 (2.7)	−.494 (−6.1)
Electrical Machinery	−.207 (−2.4)	−.112 (−3.9)	.130 (2.7)	.170 (2.1)	−.404 (−4.5)
Motor Vehicles and Equipment	.107 (1.0)	.086 (2.7)	.208 (4.1)	−.139 (−1.5)	−.044 (−0.4)
Other Transp. Equipment	−.079 (−0.7)	.030 (0.2)	.146 (3.6)	.195 (1.3)	−.532 (−2.7)
Instruments and Related products	−.147 (−1.8)	−.126 (−3.0)	.238 (3.9)	.109 (1.4)	−.398 (−2.6)
Miscellaneous Manufacturers	−.299 (−5.4)	−.106 (−2.3)	.140 (2.3)	.452 (3.4)	−.765 (−5.6)
Food and Kindred Products	−1.151 (−1.4)	−.380 (−1.9)	.052 (0.4)	1.099 (2.1)	−1.691 (−3.4)
Tobacco Products	.038 (0.3)	−.131 (−1.4)	.039 (0.3)	.581 (4.1)	−.514 (−1.9)
Textile Mill Products	.226 (0.7)	.018 (0.3)	.443 (3.3)	.151 (0.5)	−.415 (−2.7)
Apparel and Related Products	−.247 (−1.3)	−.071 (−0.9)	−.060 (−3.2)	.121 (0.6)	−.203 (−1.7)
Paper and Allied Products	−.406 (−2.2)	−.148 (−2.5)	.136 (3.2)	−.385 (−1.8)	.034 (0.3)
Printing and Publishing	−.208 (−0.8)	−.118 (−1.2)	.247 (2.3)	.392 (1.3)	−.698 (−3.2)
Chemicals and Allied Products	−.399 (−2.1)	−.245 (−2.6)	.189 (3.4)	.369 (1.4)	−.616 (−2.7)
Petroleum and Coal Products	−.781 (−1.4)	−.016 (−0.1)	.170 (2.6)	.531 (1.6)	−1.510 (−2.1)
Rubber and Misc. Plastic Products	−.226 (−2.7)	−.033 (−0.9)	.226 (3.3)	.068 (0.8)	−.485 (−4.6)
Leather and Leather Products	−.487 (−1.9)	−.075 (−1.2)	.200 (2.5)	−.234 (−1.0)	−.306 (−1.5)

Note: t-statistics in parentheses.

[a]Hours Effect is $\dfrac{pb(H-40)+pc(H-40)^2+pd(H-40)^3}{1+pa}$.

cent. This is very small relative to the 4.6 percent peak-to-trough movement in markups shown in Figure 3.

The cyclical behaviors of the components of price over marginal cost are given separately by industry in Table 5. (Again the equations are estimated by the Cochrane-Orcutt procedure, and include a constant and trends t, t^2, t^3.) Straight-time *real* wages (subtracting column 1 from column 2) are moderately procyclical in most industries. Major exceptions are the food, petroleum, and leather industries, where straight-time real wages are very procyclical, and the lum-

ber industry, where they are very countercyclical. Changes in average hours are an important procyclical component in marginal cost in almost all industries; the exceptions are the food, tobacco, and apparel industries. The cyclical behavior of productivity varies considerably across industries. Productivity is particularly countercyclical in lumber, miscellaneous manufactures, foods, tobaccos, printing, chemicals, and fuels; it is particularly procyclical in paper products, primary metals, and leather products. Price markups over marginal cost are very countercyclical in all industries except lumber and wood products, motor vehicles, and paper products. This behavior holds for durables and nondurables alike. It also does not appear to be related to an industry's average four-firm concentration ratio.[16]

V. Summary

Marginal cost is very procyclical. Using two-digit manufacturing data for after 1956, I find that a 10 percent short-run increase in production-worker employment was associated with a 2.4 percent increase in nominal marginal cost. The major cause is that employment is not perfectly flexible. In booms firms must incur a high "adjustment" cost if they expand employment, or considerable overtime pay if they expand hours per worker. Prices did not respond to the cyclical movement in marginal cost; in fact, prices were countercyclical. Markups over marginal cost decline by 3.3 percent with a 10 percent expansion. The finding of a very countercyclical markup holds across most of the two-digit industries.

This evidence is clearly inconsistent with a perfectly competitive view of manufacturing. It is also inconsistent with the view that wage stickiness is an important cause of the business cycle. Even if wage schedules are not cyclically sensitive, there is much cyclical variation in the marginal cost of labor due to variation in average hours.[17] This implies that imperfections in goods markets play a primary role in the cycle.

[16]These ratios appear in Rotemberg and Saloner (July 1984).

[17]The focus here has been on the marginal wage firms face. The results, however, also imply workers perceive a very procyclical marginal wage. This is indirect support for an equilibrium view of the labor market; and it is consistent with the conclusion Bernanke (1984) draws from disaggregate pre-World War II data.

REFERENCES

Abel, A. B., "Investment and the Value of Capital," unpublished doctoral dissertation, MIT, 1978.

Barro, R. and Grossman, H., "A General Disequilibrium Model of Income and Employment," *American Economic Review*, March 1971, *61*, 82–93.

Bernanke, B. S., "Employment, Hours, and Earnings in the Depression: An Analysis of Eight Manufacturing Industries," manuscript, Stanford Graduate School of Business, October 1984.

Berndt, E. R. and Wood, D. O., "Energy Price Changes and the Induced Revaluation of Durable Capital in U.S. Manufacturing During the OPEC Decade," MIT Energy Lab Report No. 84-003, March 1984.

Bils, M. J., "Pricing in a Customer Market," manuscript, MIT, July 1985.

Brechling, F., *Investment and Employment Decisions*, Manchester: Manchester University Press, 1975.

Carlton, D. W., "Contracts, Price Rigidity, and Market Equilibrium," *Journal of Political Economy*, October 1979, *87*, 1034–62.

Domowitz, I., Hubbard, R. G. and Peterson, B. C., "Business Cycles and the Relationship Between Concentration and Price-Cost Margins," *Rand Journal of Economics*, Spring 1986, *17*, 1–17.

Fair, R. C., *The Short-Run Demand for Workers and Hours*, Amsterdam: North-Holland, 1969.

Geary, P. T. and Kennan, J., "The Employment-Real Wage Relationship: An Inter-

national Study," *Journal of Political Economy*, August 1982, *90*, 854–71.

Gould, J. P., "Adjustment Costs in the Theory of Investment of the Firm," *Review of Economic Studies*, January 1968, *35*, 47–56.

Hall, R. E., (1980a) "Employment Fluctuations and Wage Rigidity," *Brookings Papers on Economic Activity*, 1:1980, 91–123.

______, (1980b) "Labor Supply and Aggregate Fluctuations," *Carnegie-Rochester Conference Series on Public Policy: On the State of Macro-Economics*, Spring 1980, *12*, 7–33.

Hart, R. A., *The Economics of Non-Wage Labour Costs*, London: George Allen & Unwin, 1984.

Kalecki, M., "The Determinants of Distribution of the National Income," *Econometrica*, April 1938, *6*, 97–112.

Keynes, J. M., *The General Theory of Employment, Interest, and Money*, London: Macmillan, 1936.

______, "Relative Movements of Real Wages and Output," *Economic Journal*, March 1939, *49*, 34–51.

Kydland, F. and Prescott, E., "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Nadiri, M. I. and Rosen, S., "Interrelated Factor Demand Functions," *American Economic Review*, September 1969, *59*, 457–71.

Oi, W., "Labor as a Quasifixed Factor," *Journal of Political Economy*, December 1962, *70*, 538–55.

Pigou, A. C., *Industrial Fluctuations*, London: Macmillan, 1927.

Pindyck, R. S. and Rotemberg, J. J., "Dynamic Factor Demands and the Effects of Energy Price Shocks," *American Economic Review*, December 1983, *73*, 1066–79.

Rotemberg, J. J. and Saloner, G., "A Supergame-Theoretic Model of Business Cycles and Price Wars During Booms," *American Economic Review*, June 1986, *76*, 390–407.

Sargent, T. J., "Estimation of Dynamic Labor Demand Schedules Under Rational Expectations," *Journal of Political Economy*, December 1978, *86*, 1009–44.

Shapiro, M. D., "The Dynamic Demand for Capital and Labor," *Quarterly Journal of Economics*, August 1986, *101*, 513–42.

Soderstrom, H. T., "Production and Investment Under Costs of Adjustment: A Survey," *Zeitschrift fur Nationalokonomie*, 1976, *76*, 369–88.

Stigler, G. J. and Kindahl, J. K., *The Behavior of Industrial Prices*, NBER, No. 90, New York: Columbia University Press, 1970.

Treadway, A. B., "The Multivariate Flexible Accelerator," *Econometrica*, September 1971, *39*, 845–55.

Zarnowitz, V., "Recent Work on Business Cycles in Historical Perspective: Review of Theories and Evidence," *Journal of Economic Literature*, June 1985, *23*, 523–80.

Board of Governors of the Federal Reserve System, *Federal Reserve Bulletin*, Various issues.

U.S. Chamber of Commerce, *Employee Benefits*, Economic Analysis and Study Group Biannual or Annual Surveys, 1957–82.

U.S. Department of Commerce, Bureau of Economic Analysis, *Business Conditions Digest*, Various issues.

U.S. Department of Labor, *Employment, Hours, and Earnings, 1909–84*, BLS Bulletin No. 1312–12, 1985.

U.S. Department of Labor, Special Labor Force Report No. 72, USGPO, 1965.

An Equilibrium Model with Involuntary Unemployment at Flexible, Competitive Prices and Wages

By JOHN ROBERTS*

This paper presents a general equilibrium model in which all prices and quantities transacted are explicitly chosen by economic agents: there is no Walrasian auctioneer. Multiple equilibria occur with prices and wages taking their Walrasian values. Equilibrium quantities may also be Walrasian, or they may involve some price-taking workers being rationed in selling labor. This involuntary unemployment results from self-confirming expectations of inadequate effective demand, as in some interpretations of J. M. Keynes' ideas.

The purpose of this paper is to attempt to reconcile the notion of involuntary unemployment with the hypothesis of equilibrium by constructing a closed, reasonably complete economic model which admits such unemployment as an equilibrium phenomenon.

The model in fact generates multiple equilibria which involve Walrasian, perfectly competitive prices but differing levels of economic activity. Equilibrium with full employment exists, with all agents transacting their Walrasian quantities. Simultaneously there are also equilibria at these same prices and wages in which markets fail to clear. In particular, some price- and wage-taking workers are rationed in their labor market transactions and are unable to sell as much of their labor as they desire at the given wage.[1] This involuntary unemployment arises despite the model's incorporating markets for all commodities. Further, the levels of all nominal prices and wages are endogenously determined, and in fact they are treated as choice variables of regular maximizing economic agents, as are the amounts of each good bought and sold by each agent. The agents are fully informed about one another's preferences, endowments, and production possibilities, and there is no uncertainty about any of these. Nor are markets in any sense physically separated. The agents also understand the institutions of price and quantity determination, and when making their choices they have full knowledge of any choices that have been made previously. As well, in equilibrium the agents are each acting optimally in every eventuality (not just those that would arise under some putative equilibrium behavior) while correctly forecasting both one another's choices of prices and quantities and the full implications of adopting any available course of action. In particular, no agent who is in a position to influence some prices or wages is ever mistaken about the effect of changing these, and in equilibrium no such agent finds it worthwhile, for example, to reduce wages in the face of involuntary unemployment.

It is clear that a model with these properties must depart from orthodoxy in some significant fashion. Although we assume a special structure of preferences, endowments, and technologies, under which no potential customer of a firm is also one of its potential employees, the key is in the model-

*Graduate School of Business, Stanford University, Stanford, CA 94305-5015. I would like to thank Andrew Atkeson, Jean-Pascal Benassy, Herschel Grossman, Frank Hahn, Taka Ito, Guy Laroque, Julio Rotemberg, Garth Saloner, Tom Sargent, and Nancy Stokey; the participants in seminars at CORE, LSE, Oxford, NBER, Harvard/MIT, Princeton, Georgetown, Wisconsin, USC, MSRI-Berkeley, Stanford, and Tokyo; and the referees of this *Review* for comments and suggestions. I also gratefully acknowledge the financial support of the National Science Foundation (grant nos. SES-8308723 and SES-8605666). An earlier version of this paper was circulated under the title "An Equilibrium Model with Keynesian Unemployment at Walrasian Prices."

[1] Taking this as a definition of involuntary unemployment seems to fit with much of the discussion in Chap. 2 of *The General Theory* (J. M. Keynes, 1936).

ing of the processes determining prices and individual transactions.

Neither of these processes is explicitly modeled in detail in standard equilibrium models in the tradition of Cournot and Walras. In these models, no economic agent actually sets prices. Instead, prices somehow emerge "from the market" and take whatever values are required to equate aggregate offers to buy and sell. Then, given such prices, the orders and offers that an individual has announced to the market are somehow filled, although the actual transactions that the individual makes with other agents are not generally modeled. Moreover, the results of agents' adopting nonequilibrium forms of behavior or of the implicit price adjustment and order-filling mechanisms' failing to operate are typically not specified. Thus, these models provide no formal basis for "disequilibrium" analysis.

Clearly, involuntary unemployment cannot exist in equilibrium in such models, because equilibrium involves market clearing in its very definition. If desires to buy and sell do not match, then we do not have an equilibrium and, if we assume that equilibrium will obtain (which is all we can do if we want to use these models), something will have to change. However, exactly what changes and how this occurs is outside the purview of the model; indeed, formally we cannot even say if any of the agents in the model have both the ability and the incentive to effect the requisite changes.

The model offered here is explicit about the operation of these processes, both in and out of equilibrium. Specifically, in the basic model considered in Section I, firms are treated as announcing prices for the goods they can produce and for the inputs they can use.[2] This means that there may be different prices being quoted by different firms for the same good. Knowing the announced prices, consumers place output purchase orders and input supply offers with specific firms. Actual quantities transacted are then determined by the firms' decisions of how much of these orders and offers to accept.[3]

Thus, the institutions for price and quantity determination are described by specifying which agents can act at various points, what options are open to them, what information they have when taking their actions, and what outcomes result from each set of action choices. Given preferences, endowments, and production possibilities, specifying a set of institutions in this way yields a game in extensive form. Within such a game, a (pure) strategy for an agent is a specification of the action the person will take in every possible circumstance in which he or she might have to act, not just in those that arise under some putative equilibrium mode of behavior. A well-defined outcome results from every assignment of strategies to agents, and so, given conjectures about how others are acting, each agent can determine the implications of adopting any course of behavior.

To solve this game, we search for subgame perfect equilibria (Reinhard Selten, 1975). These consist of a strategy for each agent with the property that, for each agent and for each situation in which the agent must act, adhering to the strategy and adopting the behavior specified by it is optimal for the agent, given that the other agents' current and future actions will be governed by their strategies. Equilibrium thus means first that at each decision point, each agent correctly forecasts the actions that others are currently taking and the responses from the other agents that would be elicited by the various choices the agent might make. Further, given these correct expectations, the agent chooses at each point the course of action that is optimal from that point forward. Thus, in an equilibrium with involuntary unemployment in this model, every agent correctly perceives that, taking account of the actions and reactions of the other agents, no unilateral change in behavior will benefit him or her. In par-

[2] There is only one good which does not enter production functions, and it is used as numeraire, with its price set at unity.

[3] Variants of this basic model, considered in Section II, have labor market transactions completed before output orders are placed and allow workers to set wages.

ticular, the various agents who can influence prices and wages correctly forecast that they individually will not gain by changing their prices, and the unemployed workers correctly perceive that there is nothing that anyone of them individually can do that will lead to gainful employment.

There is, of course, a huge literature that aims at generating macroeconomic inefficiency and unemployment, and it is impossible to give any satisfactory account of it here. However, the present analysis connects most clearly to a relatively few, very prominent strands of this work, and it is worthwhile mentioning these.

First is the idea of Keynesian effective demand failures based on self-confirming conjectures (see Robert Clower, 1965, and Axel Leijonhufvud, 1968): firms are unwilling to increase hiring because each forecasts that demand will be too weak to justify its increasing output, and the resultant low level of workers' incomes generates the weak demand that makes this conjecture correct. However, if all firms increased hiring together, the additional income generated could result in enough extra demand to justify the hiring. The models offered here are meant to capture formally just such notions.

Of course, these ideas have been formalized previously by Robert Barro and Herschel Grossman (1971), Jean-Pascal Benassy (1973), Jacques Drèze (1975), and others (see Benassy, 1982, for references), and aspects of their analyses reappear in the present model. Particularly important is the role of (perceived) quantity constraints and rationing in embodying the basic idea that demand constrains employment. However, in these earlier models at least some prices and wages are assumed to be fixed at levels that are not market clearing, and the models offer no basis for analysis of the opportunities for changing prices or of the incentives to do so. The present model specifically addresses these issues. Moreover, these treatments often assume that there is some unmodeled mechanism at work that ensures the maximum level of transactions consistent with voluntary exchange at the fixed prices and feasibility. Here, the determination of quantities is made explicit.

In this context, Takatoshi Ito (1979) has shown that the possibility of stochastic rationing can lead to equilibria with involuntary unemployment even when prices are exogenously fixed at their Walrasian levels. As in the Clower-Leijonhufvud interpretation of J. M. Keynes, these low-level equilibria are the result of self-confirming pessimism about demand. However, the equilibrium concept that Ito uses does not allow agents to recognize how the probability of their being rationed might depend on their announced orders and offers. Thus, unless there is a non-atomic continuum of agents, individual behavior does not appear to be fully rational in the very strong sense employed in the present paper.

A second important line of work is that dealing with macroeconomic "coordination failures" arising in non-Walrasian market settings involving search (for example, Peter Diamond, 1982; Alan Drazan, 1986) or imperfect competition (Martin Weitzman, 1982; Oliver Hart, 1982; Walter P. Heller, 1986; John Roberts, 1987; Russell Cooper and Andrew John, 1985). This literature displays the possibility of multiple equilibria and inefficiently low equilibrium levels of employment. It does not, however, generate involuntary unemployment in the sense that workers facing given prices and wages are off their supply curves in equilibrium. Instead, workers in these models transact the quantities they desire, but these are inefficiently low because imperfectly competitive restrictions on output depress labor demand and wages or because workers are monopsonistically restricting labor supply.[4] In contrast, the unemployment that arises in the present model is involuntary in precisely the sense indicated above.

A particularly striking example of coordination failures is provided by John Bryant (1983). He identifies a continuum of "rational expectations" equilibria in an economy in which different agents produce per-

[4]The Weitzman paper does obtain a sort of involuntary unemployment, but only by (in effect) fixing real wages outside the model.

fectly complementary goods. All but one of these equilibria involve inefficiently low levels of activity; the exception is one of the Walrasian allocations for the economy. The present model generates just such a continuum, and although technological complementarities play no role here, the fundamental source of the multiplicity is the same as in Bryant's work: even given prices, different agents' optimal actions are made interdependent by recognition of feasibility constraints. The inclusion of a well-specified, reasonably realistic set of market institutions in the present model enforces the point made by Bryant, and also permits the inefficiency to manifest itself as involuntary unemployment.

The work on conjectural Keynesian equilibria (Benassy, 1977; Frank Hahn, 1978; Takashi Negishi, 1979) allows both flexibility of prices and binding quantity constraints, and obtains involuntary unemployment in models incorporating price determination. However, the basis for the conjectured demand curves assumed in this work is unclear. In particular, it is not obvious that quantities would or could actually respond to price changes in the manner that the price-setting agents conjecture they will. In contrast, equilibrium in the present model requires that the conjectures about the effects of changing one's actions correspond to the actual quantities that would result if such a change were made: perceived demands and supplies must be globally, rather than just locally, correct.

Finally, work based on efficiency wage models (for example, Steven Salop, 1979; Carl Shapiro and Joseph Stiglitz, 1984; Charles Kahn and Dilip Mookherjee, 1986) yields unemployment through job rationing that is an equilibrium response to informational asymmetries. For example, a positive level of unemployment arising through wages that exceed the supply price of the amount of labor actually employed may be necessary to provide incentives not to shirk. In such circumstances, equilibrium cannot involve full employment. Moreover, the unemployment in these models may not represent any inefficiency, given the informational constraints. More generally, efficiency wage models seem best suited to explaining elements of the natural rate of unemployment. In the present model there are no problems of hidden knowledge or unobservable actions. Equilibrium may consequently be consistent both with full employment and with inefficient, "recessionary" unemployment.

The advantage of adopting the methodology used here is shown by the results that it permits. The costs are of two sorts. First, determining whether one has an equilibrium involves checking that no possible deviation is advantageous for any agent in any circumstance, and this is typically more time-consuming (although not mathematically more difficult) than verifying equilibrium conditions in a more standard model. Second, writing down an extensive form means that one has specified a very particular set of institutions. The ones assumed here do not seem to be a patently unrealistic representation of many actual markets, and they even match the informal descriptions in textbooks rather well. However, it is not immediately obvious that results proven for one set of institutions would hold under other reasonable specifications.[5]

In the next Section I describe the class of economies under consideration and the basic institutions and obtain the fundamental results. The succeeding section contains a discussion of extensions, alternative formulations, and other robustness issues, as well as suggestions for further work. The final section contains some open questions and tentative conclusions.

I. Unemployment with Perfectly Competitive Prices

Throughout we will consider a simple class of economic environments in which there are only five commodities and four types of agents, with $n \geq 2$ agents of each type.[6] The

[5]In this regard, the work of Larry Jones and Rodolfo Manuelli (1987) on a variant of the model in Roberts (forthcoming 1988) is particularly pertinent.

[6]There is no real need to assume equal numbers of each type: what is important for the present analysis is that no agent is unique. (Roberts, forthcoming 1988,

commodities are called X, Y, R, S, and M. The first pair will be outputs, the second pair will be inputs, and M will not enter the production functions. The four types of agents are labeled A, B, J, and K, and superscripts will indicate a particular agent of a given type. The first two types of agents will be called employers, producers, or firms, although they will be treated as utility-maximizing agents. The second two are called consumers or workers.

A producer of type A (respectively, B) derives utility only from the consumption of M, of which he or she holds an endowment of μ_A (resp., μ_B). Only the producers have access to the technologies of production. This may be interpreted in terms of their alone being endowed with the relevant technical know-how or some other unmarketed factor. Type A agents have the technological knowledge to permit them to produce output X from input R, while the technology available to the type B agents allows production of output Y from input S. These technologies show constant returns to scale, and we set the input-output coefficients at unity. Neither type of firm is endowed with any of X, Y, R, or S.

These assumptions will mean that any profits received will be retained by the producers as M and that supplies and demands for the other goods will not be directly affected by the levels or distribution of profits. The lack of feedbacks from profits to demand simplifies the analysis considerably.[7]

Workers of type J (respectively, K) are each endowed with ρ of R and μ_J of M, (resp., σ of S and μ_K of M) and derive utility from Y, R, and M (resp., X, S, and M). Thus, J's can supply input only to A's and can buy output from only B's. Correspondingly, K's consume the output that A's sell and supply the input used by B's. In particular, no pair of agents has a mutually advantageous trade, each consumer is a potential employee of only one type of firm and a customer of only the other, and each type of firm has only one type of consumer as potential employees and only the other as potential customers.

This separation of a firm's customers and workers is meant to model consumers' specializing in supplying labor but generalizing in consuming outputs. It has the effect of ensuring that a firm cannot directly increase the demand for its output by raising its employment or wages, because its customers' incomes are unaffected by such changes. Similarly, the supply of input (labor) to a firm is not directly dependent on its price and output levels. Some such separation or other source of leakage is important in obtaining involuntary unemployment equilibria. For further discussions of this setup, see Roberts (1987; forthcoming 1988).

We will assume that the preferences of each consumer of type J are represented by a utility function $U_J(y, \rho - r, m)$ while the consumers of type K have utility $U_K(x, \sigma - s, m)$. These functions are strictly quasiconcave, continuous, and strictly increasing for positive values, and take on values of $-\infty$ for $m < 0$.[8] The utility functions for the firms are $U_A(m)$ and $U_B(m)$, which are also strictly increasing. A specific example is the case $\rho_J = \sigma_K = \mu_J = \mu_K = \alpha > 1$, $U_K = x + \ln(\alpha - s) + \ln(m)$, and $U_J = y + \ln(\alpha - r) + \ln(m)$. In this case, the unique Walrasian, perfectly competitive prices are $p_A = p_B = w_A = w_B = \alpha$, which result in each consumer buying $(\alpha - 1)$ units of output and selling $(\alpha - 1)$ units of input (see Roberts, 1987). Of course, in this case the firms earn zero profits and so consume only their initial endowments.

Given the economic environment, we must now specify the institutions for price determination, production, and exchange. In this section, we focus on a very simple set of such institutions, while in the next section

focuses on the case of $n = 1$.) The equal numbers assumption does simplify some of the arguments, however.

[7]These feedbacks are a major source of the multiplicity of equilibria in Heller (1986).

[8]Actually, it will be sufficient that utility is a very large negative number. The point is to simplify the arguments by ensuring that consumers will not risk going bankrupt. It will become clear later that a fear of bankruptcy is not what generates the unemployment equilibria.

we consider a variety of alternatives, extensions, and enrichments.

This base set of institutions begins with each firm stating a price for its output and a wage it will pay for its input.[9] These are denominated in terms of M, the only universally desired good, which acts both as the unit of account and as the medium of exchange.[10] Workers respond to these prices and wages by announcing the amounts of input they want to sell to each firm and of output they want to buy from each. Finally, the firms decide how much of these offers and orders to accept, production is carried out, and accounts are settled.

More formally, the institutions on which we focus involve each of the $2n$ firms announcing a single price for the good it sells and a wage for the input it purchases. These announcements are made simultaneously and independently. Then, knowing these choices, $p_A^1, \dots, p_A^n, w_A^1, \dots, w_A^n, p_B^1, \dots, p_B^n$, and $w_B^1, \dots, w_B^n$, each consumer J^i of type J announces a vector $y_J^{i1}, \dots, y_J^{in}$ of output amounts and a vector $r_J^{i1}, \dots, r_J^{in}$ of input amounts and, similarly, each K^i announces $x_K^{i1}, \dots, x_K^{in}$ and $s_K^{i1}, \dots, s_K^{in}$. Again the announcements are simultaneous and independent. Each of these quantities is to be interpreted as a bona fide offer to transact the amount of the good in question with the corresponding firm at the price or wage quoted by that firm. Thus, feasibility requires $\Sigma_j r_J^{ij} \le \rho$ and $\Sigma_j s_K^{ij} \le \sigma$ for all i: workers cannot offer to sell more than their available supplies.[11] Finally, knowing the announced prices and wages and the workers' directed offers and orders, each firm A^j selects $x_A^{1j}, \dots, x_A^{nj}$ and $r_A^{1j}, \dots, r_A^{nj}$ subject to $x_A^{ij} \le x_K^{ij}$, $r_A^{ij} \le r_J^{ij}$ and $\Sigma_i x_A^{ij} \le \Sigma_i r_A^{ij}$, and similarly for each type B firm. These quantities are the amounts of each order and offer that the firm accepts, and the constraints reflect conditions of voluntary exchange and feasibility. The final allocation then involves each consumer/worker J^i buying $\Sigma_j y_B^{ij}, i = 1, \dots, n$, selling $\Sigma_j r_A^{ij}$, receiving $\Sigma_j r_A^{ij} w_A^j$ in wage payments, and paying out $\Sigma_j y_B^{ij} p_B^j$ for output, and correspondingly for each type K consumer. Meanwhile, each A^i receives $\Sigma_j p_A^i x_A^{ji}$ in revenues and pays out $\Sigma_j w_A^i r_A^{ji}$ in factor payments, and similarly for each B^i.

A fuller discussion of this institutional specification is found in Roberts (1987; forthcoming 1988). A few major points are worth noting however. First, the determination of all prices and quantities as resulting from individuals' choices is explicitly modeled: there is no unmodeled market mechanism that somehow generates prices and quantities or brings about an actual allocation from individual quantity announcements. Second, any sequence of allowable action choices leads to a clearly defined outcome. These outcomes may involve rationing, but the rationing is a result of explicitly modeled choices as well. Third, when workers make their input supply and output demand announcements, prices and wages have already been chosen. This does not yet mean, however, that workers are price takers: this will be instead a property of equilibrium. Finally, prices and wages are flexible in that each firm may set the price and wage it controls at any level it wishes.

The game induced by the environment and institutions is assumed to be common knowledge among the agents, that is, there is no private information regarding tastes or endowments, no shocks or other exogenous uncertainty, and no confusion about the options open to the agents, the outcomes resulting from any specified list of actions, or the utility values assigned to outcomes.

A (pure) strategy for an agent is a complete contingent plan specifying the action to be taken by the agent in each possible circumstance.[12] These action choices can be

[9] Thus we are assuming nondiscriminatory, linear pricing.

[10] Thus M might be interpreted as commodity money. Of course, in this context there is no need to assume that M is "used up" in consumption. It is an open question whether a dynamic version of the model can be derived that would support the presence of M, interpreted as fiat money, in the utility function.

[11] The assumption that utility is arbitrarily negative for negative holdings of M means we need not worry about consumers' planning to buy more than they can afford.

[12] We do not consider mixed strategies.

contingent on the information available to the agent at the time the person takes the action in question. Thus, a strategy for firm A^i consists first of a particular specification of p_A^i and w_A^i, and then, *for each possible vector* of price and wage announcements and of input offers and output orders, a specification of r_A^{ji} and x_A^{ji}, $j=1,\dots,n$, meeting the feasibility and voluntary exchange constraints. Similarly, a particular strategy for a consumer of type J gives his or her choice of output orders and input offers as a function of the (announced) price-wage vector. Note that one player's strategy does *not* depend on any other's strategy, because strategies per se are not observed. Instead, a strategy makes actions depend only on observables, that is, on previously chosen actions that the agent in question has observed. Note too that a well-defined outcome results from *every* specification of the strategies for each agent and not just from those corresponding to some notion of equilibrium behavior. Thus, given conjectures as to the strategies being used by the others, each agent is able to evaluate the effect on his or her utility of selecting any particular strategy or deviating from it.

An equilibrium is formally a subgame perfect Nash equilibrium of the game induced by the environment and institutions. This requires that each agent correctly forecast the strategies the others are using, and that these strategies be best responses to each other at *every* decision point, *including those that would not be reached if behavior actually is generated by the specified strategies.*

This latter requirement of sequential rationality means that if one of the agents were to deviate from his or her prescribed strategy (for example, if one of the firms were to announce a different price and wage, or one of the consumers selected different amounts of input and output, than had been expected), the agents would still find it optimal to adhere from that point forward to the behavior specified by their equilibrium strategies. In particular, this means that the firms' quantity choices must maximize profit/utility for any price and wage vector and any quantity announcements from the workers. As well, at any price-wage vector the workers must be behaving optimally, given the prices and wages and their (correct) conjectures about one another's choices and the firms' responses. This means that the workers in equilibrium must be price takers, treating the prices and wages as given and optimizing accordingly.[13] In contrast, simple Nash equilibrium would allow workers to manipulate the price and wage choices by threatening to act in a manner that would force certain price-wage choices by the firms, even though, were these choices not made, adopting the threatened behavior would not be utility maximizing. Finally, of course, each firm's price and wage choice must be the utility/profit maximizing one for it, given its correct conjecture as to the other firms' choices and given its correct forecast that the consumers' responses to the different choices it might make will be given by their equilibrium strategies.

The major focus of this paper is on showing that particular price and wage vectors and allocations of goods can be the outcome of equilibria under the specified institutions. For example, in Proposition 1 we show this for the Walrasian prices, wages, and quantities. For an economy of the type considered here, a Walrasian solution[14] is a price-wage vector (p_X, p_Y, w_R, w_S) and an allocation $\langle (x_A^i, r_A^i),\ (y_B^i, s_B^i),\ (y_J^i, r_J^i),\ (x_K^i, s_K^i),\ i=1,\dots,n\rangle$ with the following properties:

$$(1)\qquad (x_A^i, r_A^i) \text{ maximizes } U_A(\mu_A + p_X x - w_R r) \text{ subject to } x \le r,\ i=1,\dots,n;$$

$$(2)\qquad (y_B^i, s_B^i) \text{ maximizes } U_B(\mu_B + p_Y y - w_S s) \text{ subject to } y \le s,\ i=1,\dots,n;$$

[13] Of course, they do not necessarily assume that they can buy or sell as much as they want at these prices. Rather, they recognize any quantity constraints that are implied by the other agents' strategies (see below).

[14] We avoid the more common terminology of "Walrasian equilibrium" to avoid confusion. A "Walrasian equilibrium" is not an equilibrium in our meaning, although (by Proposition 1) it is the outcome resulting from an equilibrium.

(3) (y_J^i, r_J^i) maximizes

$$U_J(y, \rho_J - r, \mu_J - p_Y y + w_R r),$$
$$i = 1, \dots, n;$$

(4) (x_K^i, s_K^i) maximizes

$$U_K(x, \sigma_K - s, \mu_K - p_X x + w_S s),$$
$$i = 1, \dots, n;$$

(5) $$\sum x_A^i = \sum x_K^i, \quad \sum y_B^i = \sum y_J^i,$$
$$\sum r_A^i = \sum r_J^i, \quad \sum s_B^i = \sum s_K^i.$$

Here (1) and (2) are utility (profit) maximization for the A and B types, (3) and (4) are utility maximization for the J's and K's, and (5) is market clearing.

PROPOSITION 1: *Given a Walrasian solution, there is an equilibrium under the institutions described above such that all firms announce the Walrasian prices and wages and the resulting allocation is the Walrasian one.*

Remark: To prove this result we will not actually specify directly the full equilibrium strategies at the outset. Instead, we will follow a different route which is more enlightening and simpler. It is based upon the observation that the natural way to solve the sort of game we are considering involves adapting the method used in dynamic programming to contexts with multiple-decision makers: start with the last decision points (here, the firms' quantity choices), determine equilibrium behavior at this stage in each possible circumstance, then roll back, solving for equilibrium at the next-to-last stage using the solution at the last stage to assign outcomes to each set of choices.

Specifically, what we do is specify the choices that are to be made "along the equilibrium path." We then must ensure that no agent wants to deviate unilaterally from these specified actions. This is a trivial matter at the last stage. However, at earlier stages, when the consumers are making their quantity choices and when the firms are picking prices and wages, evaluating the attractiveness of unilateral deviations requires that we specify the behavior that follows any such deviation. Of course, this behavior must be optimal for the agents involved. Thus, we have to specify the firms' responses if any single consumer, when faced with the equilibrium prices and wages, makes quantity announcements differing from the specified ones. As well, we must also specify the quantity responses from consumers and the hiring and output decisions of the firms that follow any single firm's deviating from the specified price and wage announcement.

To complete the analysis, in situations involving multiple deviations we can make any specification of behavior so long as the strategies are optimal against one another from each point forward. In this regard, it is useful to note that for any vector of price and wage announcements from the firms, one equilibrium mode of behavior is for each consumer to announce zero and zero supply. With all other consumers ordering and offering zero, no firm to which a consumer might offer labor is receiving any orders, and so it would not hire, while no firm from which the consumer could order is offered any labor, and so it will not be able to produce to meet the expressed demand.

PROOF:

Let (p_X, p_Y, w_R, w_S), $\langle (x_A^i, r_A^i)$, (y_B^i, s_B^i), (y_J^i, r_J^i), (x_K^i, s_K^i), $i = 1, \dots, n\rangle$ be the Walrasian solution. Note that $p_X \leq w_R$, with equality if $x_A^i > 0$ for some i, and $p_Y \leq w_S$, with equality if $y_A^i > 0$ for some i. Also, $x_A^i = r_A^i$ and $y_B^i = s_B^i$, $i = 1, \dots, n$. Select $4n^2$ nonnegative numbers x^{ij}, y^{ij}, r^{ij}, s^{ij}, $i = 1, \dots, n$, $j = 1, \dots, n$, so that $\Sigma_j x^{ij} = x_K^i, \Sigma_i x^{ij} = x_A^j, \Sigma_j y^{ij} = y_J^i, \Sigma_i y^{ij} = y_B^j, \Sigma_j r^{ij} = r_J^i, \Sigma_i r^{ij} = r_A^j, \Sigma_j s^{ij} = s_K^i$, and $\Sigma_i s^{ij} = s_B^j$.

Assume each type A firm has announced $(p_A^i, w_A^i) = (p_X, w_R)$, each B has announced $(p_B^i, w_B^i) = (p_Y, w_S)$, each J^i has made offers $(r^{i1}, \dots, r^{in})$ and orders $(y^{i1}, \dots, y^{in})$, and each K^i has announced $(x^{i1}, \dots, x^{in})$ and $(s^{i1}, \dots, s^{in})$. It is clearly equilibrium behavior for the firms to hire the amounts offered and meet the demands they receive, because all feasible quantity choices for the firms yield them nonpositive profits and these particular ones yield zero. If, given the Walra-

sian prices and wages, consumer choices differ from those specified, let the firms simply maximize given the actual orders and offers. Then no consumer will gain by unilaterally deviating from the specified quantity choices because these are utility maximizing. Thus, we have equilibrium behavior given the Walrasian prices and wages.

Since the Walrasian price and wage are equal, no firm can gain by cutting its price while raising its wage. Then to ensure that no firm wishes to charge different prices and wages, it is sufficient to specify that if any firm deviates from the Walrasian price and wage except by raising its wage and lowering its price, then all offers to and orders from it are zero. In such circumstances, the workers and consumers who deal with the deviating firm when it announces the Walrasian price and wage now deal with other firms and have their Walrasian orders and offers filled.

This consumer behavior is clearly consistent with equilibrium if the deviating firm has raised its price and lowered its wage. However, if the firm has either raised both its price and wage or lowered both, matters are more complicated. If both are increased, all the workers of the relevant type would like to deal with the deviating firm but customers would prefer to buy elsewhere at the lower Walrasian price. If both the price and wage are lowered, consumers would benefit from dealing with the deviator but workers would rather find employment elsewhere at the Walrasian wage. Given the wage-price deviation, equilibrium in the continuation game clearly requires that workers and consumers effectively coordinate, with workers offering labor to the deviating firm if and only if consumers place output orders with it. However, given the deviation, either pattern can represent optimal behavior: If either group (workers or consumers) expects the other to place orders or offers only with the deviator, optimal behavior requires dealing with that firm, too, while if either group is expected not to deal with the deviator, the other should not place offers or orders with it either. The situation thus has elements of the Battle of the Sexes. To ensure that the firm has no incentive to deviate from the Walrasian price and wage, we resolve the conflict as above, that is, in favor of the workers if the price and wage are lowered, and in favor of the consumers if they are increased.

Finally, to obtain the Walrasian solution as an equilibrium outcome, we simply specify any subgame equilibrium at price-wage vectors at which more than one firm is deviating from the Walrasian levels. Note that such subgame equilibria exist; in particular, all offers and orders being zero is always one such. This follows because, with zero orders for output, no consumer can ever sell any labor and so might as well offer zero, and with zero offers, a positive output order can never be filled.

Note that although the outcome of this equilibrium is perfectly competitive, the behavior is not. Instead, each agent recognizes the quantity constraints implied by the others' choices. Of course, under the given equilibrium, these constraints are not binding. However, other equilibria exist in which they do bind, and this is the key to obtaining equilibrium unemployment at competitive prices.

PROPOSITION 2: *Given a Walrasian solution, let $0 \leq k \leq n$. Then there exists an equilibrium under the specified institutions in which all firms announce the Walrasian prices and wages and in which the equilibrium outcome gives k of the consumers of each type their Walrasian allocations and the $(n-k)$ remaining consumers of each type receive their initial endowments, buying and selling zero amounts.*

PROOF:

Let (p_X, p_Y, w_R, w_S), $\langle (x_A^i, r_A^i)$, (y_B^i, s_B^i), (y_J^i, r_J^i), (x_k^i, s_K^i), $i=1,\dots,n\rangle$ be the Walrasian solution. Note that, with strictly convex preferences, all consumers of a given type receive the same consumption bundle at the Walrasian solution, that is, $(x_K^i, s_K^i) = (x_K, s_K)$ and $(y_J^i, r_J^i) = (y_J, r_J)$ for all i. As well, all producers end up with their initial endowments.

Note that there is a Walrasian outcome in which all the firms of each type produce identical amounts. In this case, we can think of each A^i as buying only from J^i and selling only to K^i and, correspondingly, each

B^i as dealing only with K^i and J^i. We will prove the result for this case; the argument for the general case will then be clear. The argument follows the same line as that in Proposition 1.

Suppose all firms have announced the Walrasian prices and wages. Let the k agents of each type who are to be active be indexed by $1,\dots,k$. For each $i \leq k$, let J^i offer to sell r_J of R to A^i and order y_J of Y from B^i, let K^i offer s_K of S to B^i and order x_K of X from A^i, and let this pair of agents' orders from and offers to firms $j \neq i$ be zero. For $i > k$, let J^i offer zero of R and order zero of Y from each firm and, similarly, let K^i's orders and offers all be zero. Note that these transactions are feasible. Further, it is optimal for each firm to hire the labor offered to it and fill the orders it receives.

To show that this pattern of offers and orders, with the firms' hiring all the labor offered and meeting demand, is equilibrium behavior in the subgame given the Walrasian prices and wages, we must show that no consumer can gain by unilaterally deviating from the specified orders and offers, given the wages and prices and given that the firms respond optimally in the face of such behavior. That this holds for $i \leq k$ is obvious: each such J^i and K^i is realizing his or her utility-maximizing trade. For $i > k$, a firm to which worker/consumer i offers labor sees no demand for the extra output that hiring i would yield and, if active, sees no gain to hiring i in preference to its assigned worker. Similarly, a firm from which i orders does not have the input available to meet this extra demand and, if active, gains nothing from "deserting" its regular customer for i. Thus, we can specify that hiring and production remain unchanged from the pattern specified above if any J^i or K^i, $i > k$, deviates from zero orders and offers. Thus, no consumer gains by deviating.

Now we must ensure that no firm wishes to deviate from the Walrasian price and wage. The argument is essentially the same as in Proposition 1.

Again, note that wage increases coupled with a price cut cannot be profitable for any firm, so that we need consider only price increases coupled with wage decreases and movements of both the price and wage in the same direction. In the former case, suppose that A^i announces $p_A^i \geq p_X$, $w_A^i \leq w_R$, with at least one inequality being strict. Then we specify that the offers to A^i and the orders placed with it are zero. If $i > k$, so that A^i is inactive, this can be accomplished as under Walrasian prices, because zero orders from and offers to any firm are optimal for each consumer/worker if all others are ordering and offering zero. In this case, the price and wage change do not affect the firm's activity level. If $i \leq k$, so that A^i is active at $p_A^i = p_X$, $w_A^i = w_R$, then we specify that J^i shifts his or her offer of the Walrasian quantity of R and K^i shifts his or her demand for X to some firm A^j, $j \neq i$, while all other orders and offers are as at the Walrasian price and wage. The deviator A^i thus loses all its business. Further, A^j hires J^i and meets the demand from K^i, as is optimal for it. Thus, no such deviation in price and wage is profitable.

If some firm, say A^i, either lowers both its price and wage or raises both, then the conflict noted in the proof of Proposition 2 between the K's as consumers and the J's as workers arises again. In either case, we again specify that the orders and offers placed with A^i be zero. If $i > k$, this is again accomplished by having the pattern of orders and offers be as at the Walrasian price and wage. This constitutes an equilibrium in the subgame, again because zero offers to and orders from a given firm always constitute best responses to one another. Either all the J's would like to deal with A^i (if $p_A^i \geq w_A^i > w_R$), or all the K's would like to do so (if $w_A^i \leq p_A^i < p_X$), but in each case, if the one group expects the other to shun A^i's noncompetitive price and wage, so will it. Meanwhile, if $i \leq k$, again J^i and K^i coordinate in dealing with some A^j, $j \neq i$.

Finally, specifying any subgame equilibrium outcome at price-wage vectors where two firms are deviating from the Walrasian levels, we have an equilibrium with unemployment.

The key step in this argument (and in that establishing Proposition 1) is that a firm that deviates from the competitive price and wage ends up with zero activity. This was accomplished, in effect, by having the supply of

labor to any firm be infinitely elastic with respect to wage cuts below the level offered by other potential employers and by having the output demands be infinitely elastic with respect to price increases. This seems to be a natural general equilibrium analogue of Bertrand competition in both markets. Of course, other forms of behavior are conceivable, and some of these might well eliminate the involuntary unemployment of Proposition 2. What is unclear is whether any of these would support a full employment equilibrium and the Walrasian outcome in particular.

II. Alternatives and Extensions: Robustness of Results

The results in Section I indicate that a natural specification of equilibrium behavior may lead to markets' failing to clear in the context of institutions for price and quantity determination that do not seem to be an unreasonable formalization of how one might think these processes actually occur. Nevertheless, if there were natural modifications either of behavior or of the institutions that would ensure full employment, or if the failure of market clearing were crucially dependent on particular, unsatisfactory assumptions on the structure of the economy, the lack of robustness of the results would limit their interest. Thus, in this section we begin an exploration of the robustness issue.

A natural first response to the unemployment result is to ask why unemployed workers do not approach active firms and underbid the workers they are hiring.[15] However, allowing workers such an option does not obviously resolve the problem of unemployment because one would expect that the workers who are threatened with losing their jobs would respond by further undercutting. To see if this does actually eliminate unemployment, we must formalize these possibilities by specifying explicitly institutions that include them.

One natural place to insert the possibility of workers' offering to work for less is after the firms have made their quantity decisions. Suppose then we allow each worker at this stage to approach some firm and offer to work at a lower wage than the firm had earlier quoted, but that, if a firm is approached in this fashion, the worker(s) it was going to hire can make a counteroffer. The firm then will have the option of accepting either, both, or neither of these offers. Our object is to show that the unemployment equilibria described in Proposition 2 survive this alteration in the institutions.

To complete the institutional specification, the process for determination of quantities after such underbidding must be given. There are a wide variety of options here, both with regard to the amount of labor to be provided when undercutting occurs and the amounts of output to be supplied.

On the input side, the simplest solution is to require that, if underbidding occurs, the workers doing so must offer to supply the same amount of labor that the firm had originally planned to hire, even though this amount might exceed the workers' competitive supply at the lower wage. Given the orders it faces, the firm will then hire the same amount as before. Note that if the workers offered some different quantity, the firm might change its output level, and the rational forecast of this would alter offers and orders throughout the economy.

Given this requirement on input supplies, the simplest rule on the output side is that workers do not get to revise their quantity orders from those made earlier. Note, however, that if underbidding does occur in equilibrium, then this will be forecast and the orders placed will allow for the reduced wages and the hiring that results.

In this context, it is easy to see that the possibility of undercutting may have no effect. In particular, consider the case in which the workers' endowments are $\mu_J = \mu_K = \rho_J = \sigma_K = \alpha$ and the utility functions are $U_J(z, t, m) = U_K(z, t, m) = z + \ln(\alpha - t) + \ln m$. In this case, the Walrasian quantities are $(\alpha - 1)$ and the price and wage levels are

[15] Recall that we have already dealt with firms' cutting wages, although we have not allowed discriminatory wage offers. However, it seems clear that even discrimination would not upset the unemployment outcome.

α. Then, if an agent of type J, say, has initially ordered $y=0$ and has sold no labor, the lowest wage he or she would accept if he or she must provide $(\alpha-1)$ units of labor and still get no output is $\underline{w}=\alpha$, that is, the prevailing Walrasian wage. Thus, such an agent in fact will not underbid the Walrasian wage, and the equilibrium described earlier survives, even with no counteroffers from the employed.[16]

Of course, this specification may not seem a fair representation of the idea of undercutting, because the only reward to working is in the increase in M and not in increased consumption of output. At the opposite extreme, we could allow all agents to revise their quantity decisions after any underbidding, but this would lead to a highly complicated model. A simpler alternative is that if one worker ends up supplanting another through underbidding, then the former simply takes over (and pays for) the output order the other had previously placed.

In this case, suppose the strategies for the initial stages are as in Proposition 2, so that the prices and wages quoted by each firm are equal, the quantities bought and sold by agents $1,\dots,k$ are the Walrasian amounts and the quantities transacted by the other agents are zero. Clearly no employed agent can gain by underbidding. Consider then an unemployed agent, say J^i, $i>k$. Since no firm A^j, $j>k$, sees demand for its product, it is pointless to offer to work for such a firm. Let J^i then approach some active firm, A^j, $j<k$, and offer to work for w^i, which is less than the Walrasian wage, w^W. Letting r^W and y^W denote the Walrasian quantities of R and Y and p^W the Walrasian price of Y, the lowest value that i would be willing to set for w^i is given by $U_J(y^W, \rho-r^W, \mu_J-p^W y^W+\underline{w}r^W)=U_J(0,\rho,\mu_J)$. This same value also defines the lowest counteroffer that J^j, the worker tentatively assigned to A^j, would make.

It is easy to see that equilibrium requires that A^j accept the lower of the wages offered by J^i and J^j. Then, the logic of Bertrand competition means that the unique equilibrium in the subgame is for both J^i and J^j to bid $\underline{w}$. But then, even if hired, J^i gains nothing over being unemployed, and so we may specify that the person never bothers to underbid in the first place. Thus, the allocation from Proposition 2 remains an equilibrium outcome, even when underbidding is permitted.[17]

Another approach to allowing workers to influence wages is simply to have the firms at the first stage announce only output prices, while each worker announces the wage at which the person is willing to sell his or her labor. It would then be natural to have each firm announce an amount of labor it wants to hire from each worker and, simultaneously, each consumer announce an amount of output he or she wants to buy from each firm. Finally, workers could decide how much of each employment offer to accept, and firms could then decide on how much to produce and sell.

Note that a firm anticipating zero orders will make no offers of employment, and if consumers expect a firm not to hire there is no sense placing orders with it. Thus the arguments used earlier can be adapted to obtain a result paralleling Proposition 2 for these institutions.

A second possible alteration in the setup involves agents' behavior within the original institutions. The unemployment equilibria in Proposition 2 involve unemployed workers' ordering zero quantities of output and offering zero quantities of labor. This is rational (that is, equilibrium) behavior; no individual can unilaterally change orders and offers and be better off as a result. Still, these orders and offers may well be less than the workers' "effective demands," defined as maximum

[16]This argument will clearly work so long as $U(0,\tau-t,\mu+wt)\le U(0,\tau,\mu)$, where τ and μ are the endowments and t and w are the Walrasian labor supply and wage, that is, as long as $U_2/U_3>w$ at $(0,\tau-t,\mu+wt)$.

[17]This argument obviously pushes the Bertrand logic very hard. Nevertheless, it does not seem totally improbable to suppose that an unemployed worker might not expect to be able to supplant an employed one by offering to work for less because the latter will be willing to match his or her offer.

amounts that the individual would be willing to transact on each market, given that the person forecasts that the transactions on the other market will be constrained to zero (Benassy, 1973).

Benassy (1982) has argued that, in such a situation, agents should announce their effective demands, rather than their anticipated transactions, because the latter too easily lead to unemployment equilibria. Realizing beneficial trades requires that agents somehow transmit the information that they are willing to trade, and announcing anticipated transactions may fail to do this. Moreover, in the present context it would seem costless to increase one's orders and offers, given the behavior of the others, while if others also did so, the unemployment might be broken.

While Benassy's effective demands have been questioned on various grounds (Ito, 1979; Gale, 1983; Roberts, forthcoming 1988), we wish to show that, even if agents do announce effective demands, equilibrium may still fail to yield market clearing.

Consider again the particular utility functions and endowments discussed above and recall that the Walrasian prices and wages are all equal to α. Now consider the effective demands of a worker at these prices and wages who anticipates making zero transactions. For output, these are given by the solution of

$$\max_{z}\; z+\ln\alpha+\ln(\alpha-\alpha z),$$

which is uniquely $z=0$, while the effective supply of input is the solution of

$$\max_{z}\; \ln(\alpha-z)+\ln(\alpha+\alpha z),$$

which is $(\alpha-1)/2$.

Thus, in this case, if workers announce their effective demands for input and output, they actually do indicate that they are willing to work at the going wage. However, demands are still zero, so hiring is zero. From this, we see that unemployment of the type demonstrated in Proposition 2 still may persist even if workers state their effective demands.[18]

It may be worth noting that, if workers were all to announce their Walrasian supplies and demands, then full employment would be achieved (this is just Proposition 1), but this behavior is, in a sense, risky. In the context of the example, suppose one individual announces his or her Walrasian demand and supply but the "corresponding" worker-consumer of the other type announces a positive labor supply and a zero output demand. Then the first agent can end up with negative money holdings, because the person is committed to buying output but sells no labor. This argument suggests that out unemployment equilibria would not be eliminated by requiring "trembling hand" perfect equilibria (Selten, 1975) rather than just subgame perfect ones.

While the preceding results formalize the "Keynesian" intuition that expectations of low levels of demand can be self-confirming and can result in equilibria with involuntary unemployment, they do not exactly match with earlier discussions of these ideas (Clower, 1965; Leijonhuvfud, 1968). In that literature, it was the expectations held by each firm about the demand that would be forthcoming at given prices as it varied its hiring that were central; here firms already have received output orders when making their hiring decisions and so need make no such forecasts. Instead, workers are attempting to predict one another's quantity choices, and the inefficient equilibria arise from their pessimism.

However, a modification of the assumed institutions yields a more central role for firms' beliefs and captures much of the standard view while still maintaining the possibility of unemployment equilibria.

[18]Again it is clear that this argument would work more generally and, in particular, if either the effective demand for output when constrained to sell no labor or the effective supply of labor when constrained to buy no output is zero. For this it is sufficient that either $U_2/U_3 > w$ or $U_1/U_3 < p$ at (o,τ,μ).

Specifically, suppose that, as before, each firm first announces a price and wage. Next, however, workers make (directed) announcements not of both their labor supply and output demand but, instead, just of amounts of labor they wish to sell. Then, knowing the workers' labor offers, each firm independently makes its hiring decisions, which consist, in each case, of how much of the offers made to it to accept. At this point, the firms become committed to paying the contracted amounts.[19] Next, the workers place output orders, knowing the amount of labor they have sold and the wages they will receive. Finally, firms decide how much of these to fill. Assume, as before, that at each stage all agents are fully informed about all previously made choices.

An equilibrium under these institutions is again a subgame perfect equilibrium in the induced game, and again we find such equilibria by examining the stages sequentially, beginning with the last.

The determination of output levels at the last stage is completely trivial, because the firms' costs are, at this stage, fixed. Thus, we can assume that each will always produce the smaller of the total ordered from it and its maximum feasible output (as determined by its hiring). Note, however, that if production is smaller than orders, the firm will have to decide how output is to be allocated among those ordering from it. Of course, the firm is indifferent over all such distributions of the given output, so any pattern of such rationing is consistent with equilibrium in these subgames.[20]

In equilibrium, consumers will correctly foresee the allocation of output resulting from any specified list of output orders. Using this forecast and knowing prices, wages, and hiring, each consumer announces an output order to maximize his or her utility, given conjectures about the offers others are placing.[21] Equilibrium then requires that these conjectures are correct, so that each consumer's order is optimal against the orders that others are actually placing.

A particular specification of equilibrium at this stage then allows each firm to decide how much to hire by examining the sales that result from various hiring levels, given its conjectures about the other firms' hiring. Again, subgame equilibrium is characterized by each firm's choice being optimal, given correct conjectures about the others' decisions and correct forecasts regarding the resultant orders and sales. Once more rolling back, workers then can forecast the full implications of making different patterns of labor supply offers. Solving for a subgame equilibrium here then gives the relations between prices, wages, and quantities that firms use in deciding on their price and wage announcements.

The sequential determination of inputs and outputs has important consequences here. Not only are workers sure of their incomes when they place their output orders, but also they know the supplies potentially available from the various firms. This can, in some circumstances, break the zero-activity level equilibria that involve each consumer ordering zero because the person expects no labor to be supplied and thus no output to be available, and simultaneously offering no labor because he or she expects zero output demand and hiring. Suppose now a firm has announced a price at which demand is positive even when its customers cannot sell any labor. The firm can then anticipate that if it hires it will receive demand. Consequently, workers know that if the firm's wage is lower than this price, it will in fact do some hiring if it is offered labor. They will then make positive labor supply offers so long as doing

[19] Feasibility may thus require that the firms' hiring be constrained by their endowments of M.

[20] Note that if orders are smaller than hiring, the firm need not use all the labor for which it has contracted to pay. In the sequel we will assume, for simplicity, that if a firm has hired excess labor, the full amount must be supplied by workers and is unavailable for their own consumption. Of course, equilibrium will involve hiring equal to sales.

[21] Existence of an optimal order may depend on the rationing rule: see Bénassy (1982).

so is optimal given the firm's wage, the prices they face for output, and their forecasts of output availability.

Nevertheless, equilibria parallel to those in Proposition 2 may still exist with these institutions.

PROPOSITION 3: *Consider the economy with endowments* $\mu_J = \mu_K = \rho_J = \sigma_K = \alpha$, $\alpha > 1$, *and consumer preferences* $U_J(z, \alpha - t, m) = U_K(z, \alpha - t, m) = z + \ln(\alpha - t) + \ln m$. *Let the institutions be specified as above with sequential supply offers and output orders. Then for each* k, $0 \leq k \leq n$, *there exists an equilibrium in which all firms announce the Walrasian price and wage levels, consumers* $1, \dots, k$ *of each type transact their Walrasian quantities and consumers* $k+1, \dots, n$ *transact zero amounts and consume their initial endowments.*

PROOF:

As in Proposition 2, let the agents of each type who are to be active be indexed as $1, \dots, k$, while the unemployed are indexed by $k+1, \dots, n$.

We again limit ourselves to describing the actions chosen along the equilibrium path and to ensuring that there are never incentives to deviate. These actions are: all firms announce prices and wages equal to the Walrasian level, α; consumer i of type J (respectively, K) offers $(\alpha - 1)$ of R to A^i (resp., of S to B^i) if $i \leq k$ and offers an amount less than or equal to $(\alpha - 1)/2$ of R to A^i (resp., of S to B^i) if $i \geq k$; each firm i hires the full amount offered to it if $i \leq k$ and hires 0 otherwise; consumer J^i orders $(\alpha - 1)$ of Y from B^i (resp., K^i orders $(\alpha - 1)$ of X from A^i) and orders zero from all other firms if $i \leq k$, while, for $i > k$, J^i and K^i order zero from all firms; and, finally, each firm fills its orders.

Clearly, the firms' behavior at the last stage is optimal, as is that of the consumers $1, \dots, k$ of each type at the order stage. As well, because (as noted earlier) the optimal purchase of output at a price of α is zero if one has sold no input, the consumers with $i > k$ are also acting optimally in their orders.

The hiring of the firms yields each of them zero profits, and no deviation from this behavior can generate positive profits with prices and wages set at the Walrasian levels. Thus, the specified hiring behavior is consistent with equilibrium.

Obviously, given the prices and wages, no active consumer can benefit from changing his or her labor offer. If an inactive worker changes either the size of the offer or the firm to which it is made, then we specify that the person still sells zero. That this is consistent with equilibrium is simply seen by noting that if the firm hires any positive amount from this worker, then it earns zero if it manages to sell the resultant output (and so does not gain) and otherwise loses money.

Thus we arrive at the stage where prices and wages are set.

Consider first an active firm that considers deviating from the Walrasian price and wage. If this deviation involves a wage cut, then we simply specify that the firm receive no labor offers. (Any worker who would have offered labor to this firm now goes to some other, nondeviating firm and offers it his or her Walrasian supply. This is accepted, and any customer who would have bought from the deviator along the equilibrium path now deals with the firm with the increased hiring.) Thus, no wage cut helps.

Suppose then the deviator, say firm A^i, raises its wage. Of necessity, it will have to raise its price even more to gain from this deviation. Let its choices then be $p > w > \alpha$. We must now ensure that this gets zero sales. Note that we cannot simply adopt the technique used in Propositions 1 and 2 of having the workers expect the demand all to go to firms charging the low, Walrasian price, because here workers move first. Thus, if all of them go to the high-wage deviator, it alone will have output and customers will have no effective option but to deal with this firm. Nevertheless, it still turns out to be an equilibrium for workers to shun the high-wage deviator, in essence because they expect it will not, in fact, hire them because it correctly forecasts a zero demand for its product.

The specified response to an A setting $p > w > \alpha$ is then for all workers to make Walrasian offers to nondeviating firms, for these to be accepted, for consumers each to

order their Walrasian quantities, placing the orders in such a way that they can be met, and for the nondeviators to fill the orders. In particular, if the deviator is A^j, let J^i and K^i, $i \neq j$, deal with A^i and B^i, while J^j and K^j deal with A^{j+1} and B^{j+1} if $j < n$ and with A^1 and B^1 otherwise.[22]

The key issue is whether any worker-consumer can gain by deviating from this pattern. Clearly none can gain by changing his or her order, given the specified pattern of labor offers and hires, nor can any K gain from changing his or her offer of S. But can a worker of type J offer labor to the deviating, high price and wage A and get more than the Walrasian utility level?

Suppose first that the deviator is an active firm, say A^1, and that J^1 offers labor to A^1 rather than A^2. (The case in which J^j, $j \neq 1$, offers labor to A^1 is identical.) The other workers must continue the offers specified above because their offers are made simultaneously with J^1's, but the firms will be able to adjust their hiring in response to this deviation and the consumers will be able to alter their orders.

Suppose then that A^1 hires r^1 from J^1. Let A^i continue to hire $(\alpha-1)$ from J^i, $i \geq 2$, let B^2 continue to hire $(\alpha-1)$ from K^2 and zero from K^1, and let B^i hire $(\alpha-1)$ from K^i, $i > 2$. (Recall that, following the deviation by A^1 but absent the subsequent deviation by J^1, K^1 would be employed by B^2. Also recall that B^1 receives no offer of labor when A^1 deviates.) Now, let consumers J^i (resp., K^i) order their optimizing quantities from B^i (resp., A^i), $i \geq 2$, let K^1 order his optimal quantity, which, because he or she is not hired, is zero, and suppose that if J^1 places an order with any of the firms $B^1,\ldots,B^n$, it is not filled. Note that this behavior for the firms is optimal because B^1 has no output to offer and each B^i, $i \geq 2$, is selling its entire output and is indifferent about who receives it. Then the best that J^1 can do is receive 0 output. Further, since K^1 is not hired, firm A^1 will get no demand and so will actually not accept J^1's offer. Thus, J^1 ends up with his or her initial endowment and is worse off for deviating. This in turn means no active firm has an incentive to deviate from the Walrasian price and wage.

Finally, a similar argument shows that A^i, $i > k$, does not gain by raising its price and wage because, again, no worker-consumer will optimally deal with it in preference to making his or her Walrasian transaction with another firm.

Thus we have the possibility of some fraction of the economy enjoying its Walrasian consumption vectors while the remainder of the agents end up in equilibrium unable to transact at all, even though these unemployed workers explicitly state their willingness to work at currently offered wages and are in all respects identical to the employed. Moreover, the reason they are not hired is that firms simply do not forecast sufficient demand. At the same time, there is also an equilibrium at the same Walrasian prices and wages with full employment. However, the decentralized market institutions do not effectively coordinate individual actions and ensure that the efficient equilibrium is reached.

There are many obvious directions for further extensions of this work. We mention only four of these.

First, a central feature of the proof of the existence of unemployment equilibrium in Proposition 3 is that full employment obtains if one of the firms raises its price and wage above the Walrasian level. (Note that this is not the case in Proposition 2, nor is this pattern employed in Proposition 3 if the firm makes the possibly more intuitive deviation of cutting its wage offer in the presence of unemployment. In these cases, a deviation simply results in the firm's losing business, as under Bertrand competition, but not in a change in the aggregate level of activity.) Of course, the firm itself does not gain from this sort of deviation, so unemployment remains as an equilibrium phenomenon. However, it would clearly be interesting to see if unemployment can be sustained under sequenced orders and offers without there being higher

[22] Note that we thus have full employment "off the equilibrium path" if a firm deviates in its price and wage choice by raising its wage. We will return to this below.

levels of employment off the equilibrium path than along it.

More generally, it can be argued that the responses to a wage increase assumed in Proposition 3 involve too great an economy-wide reaction to a single firm's actions: recognition of observability questions and the need for search in real economies argue for some measure of continuity. There is certainly merit to these points, and it would be very interesting to investigate whether continuity can be imposed while maintaining the existence of both full employment and unemployment equilibria. Results of this sort are given an example of a world of monopolies (that is, $n=1$) in Roberts (forthcoming 1988), but none are yet available for the competitive environments considered here.

A further feature of this proof (which also was important in Propositions 1 and 2) was the heavy use made of the firms' technologies' showing constant returns to scale. This implied both that the firms earned zero profits in competitive equilibrium and that they were willing and able to expand to absorb the workers and customers of a competitor at the given price and wage. If nonconstant returns are assumed, firms will no longer be indifferent regarding output levels at the Walrasian prices and wages. In this case, many aspects of the arguments made above would fail. Clearly, it would be interesting to determine whether equilibrium under decreasing returns is still consistent with both full employment and unemployment at Walrasian prices and to study the nature of equilibrium with increasing returns.

In a related vein, one should relax the special structure of endowments and preferences assumed here. It seems that some separation of a firm's potential employees and potential customers is important for coordination failures, but how much such separation is needed? A more general model might allow workers to be interested in a wide variety of outputs, so that their demand would be a small fraction of the total demand facing their employer. It seems likely that results paralleling those obtained here would then hold for "approximate equilibria" in which each firm ignores the impact of its hiring and wage payments on its demand, but it is unclear if this would be the case for full equilibria. As well, it would be worthwhile to allow workers to be able to supply labor to several different industries. As long as they continued to be uninterested in consuming any output to whose production they could contribute, this should lead to no alteration in the results.

The fourth desirable extension would be to incorporate the possibility of government which could try to use planning or fiscal policies to lead the economy to efficient equilibria. More generally, one would like to investigate whether there are other alterations in the institutional framework that seems natural and would yield more efficient outcomes.

III. Conclusions

The question of whether involuntary unemployment can be an equilibrium phenomenon has troubled economists for at least 50 years. If we use standard models in the tradition of Cournot and Walras, the answer must be that equilibrium implies market clearing. But the process by which prices and quantities are determined is not fully specified in such models, and, in effect, market clearing is made part of the definition of equilibrium.

In this paper I have attempted to be fully explicit about these processes and then have shown that the natural concept of equilibrium under reasonable institutions need not imply market clearing: absent a Walrasian auctioneer or some similar institution, the decentralized decisions of individuals may be left inefficiently uncoordinated by market forces, even though all agents are very well informed, have fully rational expectations, and maximize accordingly, and even though prices and wages are flexible.

The analysis presented here points to a crucial role for expectations or "animal spirits." This role is very much in accord with Clower's treatments of Keynesian ideas and with later developments of his ideas by others concerned with macroeconomic issues. Thus, I have focused the discussion of equilibrium without market clearing on involuntary unemployment. There is still, however, a serious question of whether the present analysis has any real power in helping us

understand actual unemployment, business cycle activity, and the like. The problem is that the model generates too many equilibria: we have the range of equilibria at Walrasian prices identified in Propositions 2 and 3, and it is possible that there might be others as well.[23] While a multiplicity of equilibria seems central to any formalization of Keynesian ideas, the extreme multiplicity present here deprives the model of predictive value.[24]

Nevertheless, the ease with which we have obtained equilibrium without market clearing, once the institutions for price and quantity determination are explicitly modeled, seems to me to raise important questions. Is it simply that the institutional specification given here is the wrong one, and that market clearing would result automatically if the right institutions were assumed? A centralized Walrasian market would have this property, but assuming these institutions actually exist seems ridiculous. Or, perhaps, is there some equilibrium selection mechanism that leads markets to the most efficient equilibria but that has been ignored here? The fixed-price literature in effect seems to assume something like this in positing that, given prices, the rationing is efficient, but the forces that would bring this about are not obvious. Or, perhaps, have we simply been too hasty in assuming that our models with built-in market clearing will accurately predict actual market outcomes?

[23]In particular, it is an open question whether there can be equilibria in which firms announcing prices and wages different from the Walrasian ones are active. The logic of Bertrand equilibrium might suggest that there ought not to be such equilibria, but the issue seems complex.

[24]Note that there is still the possibility that the presence of autonomous government demand might yield more clear-cut results.

REFERENCES

Barro, Robert J. and Grossman, Herschel I., "A General Disequilibrium Model of Income and Employment," *American Economic Review*, March 1971, *61*, 82–93.

Benassy, Jean-Pascal, *Disequilibrium Theory*, doctoral dissertation, University of California, Berkeley, 1973.

______, "A Neo-Keynesian Model of Price and Quantity Determination in Equilibrium," in *Equilibrium and Disequilibrium in Economic Theory*, G. Schwodiauer, ed., Dordrecht: D. Reidel Publishing, 1977, 511–44.

______, *The Economics of Market Disequilibrium*, New York: Academic Press, 1982.

Bryant, John, "A Simple Rational Expectations Keynes-Type Model," *Quarterly Journal of Economics*, August 1983, *98*, 525–28.

Clower, Robert W., "The Keynesian Counterrevolution: A Theoretical Appraisal," in *The Theory of Interest Rates*, Frank H. Hahn and Frank P. R. Brechling, eds., London: Macmillan, 1965.

Cooper, Russell and John, Andrew, "Coordinating Coordination Failures in Keynesian Models," Cowles Foundation Discussion Paper 745, Yale University, 1985.

Diamond, Peter, "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, *90*, 881–94.

Drazan, Alan, "Involuntary Unemployment and Aggregate Demand in an Optimal Search Model," Working Paper No. 32-86, Foerder Institute for Economic Research, University of Tel Aviv, 1986.

Drèze, Jacques, "Existence of an Exchange Equilibrium Under Price Rigidities," *International Economic Review*, June 1975, *16*, 301–20.

Gale, Douglas, *Money: In Disequilibrium*, New York: Cambridge University Press, 1983.

Hahn, Frank H., "On Non-Walrasian Equilibria," *Review of Economic Studies*, February 1978, *45*, 1–18.

Hart, Oliver, "A Model of Imperfect Competition with Keynesian Features," *Quarterly Journal of Economics*, February 1982, *97*, 109–38.

Heller, Walter P., "Coordination Failure under Complete Markets with Applications to Effective Demand," in *Equilibrium Analysis: Essays in Honor of Kenneth J. Arrow*: Vol. II, Walter P. Heller, Ross M. Starr, and David A. Starrett, eds., Cambridge: Cambridge University Press, 1986, 155–75.

Ito, Takatoshi, "An Example of a Non-Walrasian Equilibrium with Stochastic Rationing at the Walrasian Equilibrium Prices," *Economics Letters*, *2*, 1979, 13–19.

Jones, Larry and Manuelli, Rodolfo, "The Coordination Problem and Equilibrium Theories of Recessions," Kellogg School of Management, Northwestern University, June 1987.

Kahn, Charles and Mookherjee, Dilip, "Competitive Efficiency Wage Models with Keynesian Features," Research Paper No. 901, Graduate School of Business, Stanford University, 1986.

Keynes, John Maynard, *The General Theory of Employment, Interest and Money*, London: Macmillan, 1936.

Leijonhufvud, Axel, *Keynesian Economics and the Economics of Keynes*, London: Oxford University Press, 1968.

Negishi, Takashi, *Microeconomic Foundations of Keynesian Macroeconomics*, Amsterdam: North-Holland, 1979.

Roberts, John, "General Equilibrium Analysis of Imperfect Competition: An Illustrative Example," in *Arrow and the Ascent of Modern Economic Theory*, George Feiwel, ed., London: Macmillan, 1987, 415–38.

______, "Involuntary Unemployment and Imperfect Competition: A Game Theoretic Macro Model," in *The Economics of Imperfect Competition and Employment: Joan Robinson and Beyond*, George Feiwel, ed., London: Macmillan, forthcoming 1988.

Salop, Steven, "A Model of the Natural Rate of Unemployment," *American Economic Review*, March 1979, *69*, 117–25.

Selten, Reinhard, "A Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, *4*, 1975, 25–55.

Shapiro, Carl and Stiglitz, Joseph E., "Unemployment Equilibrium as a Worker Discipline Device," *American Economic Review*, June 1984, *74*, 433–44.

Weitzman, Martin, "Increasing Returns and the Foundations of Unemployment Theory," *Economic Journal*, December 1982, *92*, 787–804.

Ski-Lift Pricing, with Applications to Labor and Other Markets

By ROBERT J. BARRO AND PAUL M. ROMER*

The market for ski runs or amusement rides often features admission tickets, no explicit prices for rides, and queues. Although the prices of admission tickets are much less responsive than the length of queues to variations in demand, we show that the outcomes are nearly efficient under plausible conditions. Then we show that similar results obtain for some familiar congestion problems and for profit-sharing schemes in the labor market.

During Christmas or spring vacation, most ski areas have long lines. The same is true for Disneyland and other amusement parks in peak season. This type of crowding does not depend on surprises in demand, but instead is systematic. Most economists look at chronic queuing and conjecture that the suppliers would do better by raising prices. Further, most economists would argue that the failure to price properly leads to inefficient allocations of rides, as well as improper investment decisions. But the regular occurrence of lines in some markets suggests that it is economists, rather than suppliers (who have survived), who are missing something.

We argue in this paper that competitive suppliers of ski-lift services (amusement rides, etc.) may rationally set prices so that queues occur regularly and are longer at peak times. Under plausible assumptions, this method of pricing can support efficient allocative decisions. In equilibrium, owners of ski areas set prices for all-day lift tickets (or equivalently, for admission tickets to amusement parks) by maximizing profits subject to a downward-sloping demand curve. This appearance of monopoly power leads to no inefficiency. Moreover, the equilibrium price charged for a lift ticket may not rise with expansions of demand. Sticky prices may be consistent with optimization by suppliers and with efficient choices of quantities.

There are two distinct effects in operation under lift-ticket pricing, either of which is sufficient to imply that setting the marginal cost of a ride equal to zero does not lead to distortions. The first effect, which we call the package-deal effect, arises under any two-part pricing scheme with quantity constraints. Someone who buys 10 units of a good at $1 each from a local shopkeeper will be indifferent between standard pricing at $1 per unit and a two-part pricing scheme with a $10 entry fee, a per unit price equal to zero, and a limit of 10 units per customer.[1]

The second effect, which we label the homogeneity effect, describes conditions under which congestion leads to no efficiency losses. The usual argument, dating back to the two-roads problem of Frank Knight (1924), is that free entry into one of two activities equates average rather than marginal returns, and thereby leads to welfare losses. This conclusion does not follow if the same quantity equates average and marginal returns. In the examples we consider, this coincidence arises under natural assumptions about how congestion affects individual utility. For example, suppose that total output in activity i takes the form $x_i f(n_i)$, where x_i is a measure of capacity or desirability and n_i is the number of people who participate. For activities 1 and 2, the values of n_1 and n_2 that equate the average prod-

*Department of Economics, Harvard University, Cambridge, MA 02138, and University of Rochester, Rochester, NY 14627, respectively. We are grateful for support of research from the National Science Foundation. We have benefited from comments by Gary Becker, John Boyd, Kevin Murphy, and Sherwin Rosen.

[1]This point is familiar in the context of the labor market. See Robert Barro (1977) and Robert Hall (1980).

uct, $x_1 f(n_1)/n_1 = x_2 f(n_2)/n_2$, also equate the marginal product, $x_1 f'(n_1) = x_2 f'(n_2)$, if $f(n_i)$ is homogeneous of some degree. The same holds true if the dependence on x and n takes the form $g(x, n)$, where g is homogeneous of degree 1.

In the case of ski area or amusement park pricing, we make an additional, subsidiary point. Queues may have an effect on the allocation of resources that has nothing to do with the cost of time. The package-deal effect applies only if there are quantity constraints. Ski areas and amusement parks place no explicit constraints on the quantities that each person can consume, but the queues impose an implicit constraint. If there are x total rides available and n people show up to consume these rides, the queue may act purely as a symmetric allocation mechanism to provide x/n rides to each person. To emphasize this alternative view of the role played by queues, we make the extreme assumption that time spent at a ski area or amusement park is inherently valuable so that the cost of time spent in the queue is approximately zero. This assumption may be inappropriate for other settings, but it is inessential for the operation of either of the two basic effects. The homogeneity effect does not depend on any quantity constraints; the package-deal effect does require them, but these constraints can arise in ways other than queues.

Because ski-lift pricing illustrates both effects as well as the subtle fashion in which quantity constraints can arise, we start with this example and analyze it in detail. Section I considers the simplest case, where ski areas are identical and consumers differ only in the cost of going skiing. We compare the alternative modes of pricing, with emphasis on the assumption concerning the cost of time. Then we show why lift-ticket prices may not vary with changes in demand. Section II considers extensions where individuals differ in preferences, and ski areas differ in characteristics. Section III presents applications of the package-deal and homogeneity effects, including the two-roads problem, a common-property fishing problem, and profit-sharing arrangements for employees of a firm.

I. The Supply and Demand for Ski-Lift Services

To highlight the operation of the package-deal effect, we analyze first the equilibrium with conventional pricing for individual rides at a ski area, and then show how this equilibrium can be repackaged into one with an entry fee—the lift ticket—and a price per ride set equal to zero. Section I, Part A, describes the market for rides on a ski lift and works out the equilibrium. Section I, Part B, illustrates how the quantities and prices from this efficient solution can be replicated in an equilibrium with lift-ticket pricing and queues. Section I, Part C, discusses how the results change when we allow for the opportunity cost of time spent in a queue, then considers the factors that might influence the choice between the alternative pricing arrangements. Section I, Part D, illustrates why lift-ticket prices may not respond to changes in demand. The analysis in Section I assumes that ski areas are identical and that individuals differ only in a limited sense. In this context only the package-deal effect is present. Section II discusses the homogeneity effect, which arises when we allow for differences among individuals and ski areas.

A. *Equilibrium with Ride Tickets*

Consider a group of identical, competitive ski-lift operators, each of whom sells ride tickets at a price P per ride. Each firm has a fixed capacity and therefore supplies inelastically the total quantity of rides x. Flexibility in this quantity at some positive marginal cost is more realistic, since suppliers can open more lift lines or perhaps operate the existing ones at greater speed. But these modifications do not change our results. In the present case the industry's total supply of rides is Jx, where J is the fixed number of firms.

Let q_i be the quantity of ski rides for the ith person. This individual chooses q_i to maximize

$$U^i = U^i(q_i, z_i)$$

subject to

$$Y_i = Pq_i + z_i + c_i,$$

where Y_i is real income, z_i is goods other than skiing (price normalized to one), and c_i is an individual-specific, lump-sum cost of going skiing. This cost is quasi fixed because it depends only on the decision to go skiing, not on the number of ski runs consumed.

For those who ski ($q_i > 0$), the determination of q_i can be described in the usual way by a downward-sloping, income-compensated demand curve, $q_i = D^i(P)$. In this section we neglect variations across individuals in these demand functions (contingent on participation in skiing)—that is, we assume $q_i = D(P)$. A later section allows for heterogeneity of demands, which could reflect differences across the population of skiers in tastes for skiing or in net incomes, $Y_i - c_i$. Given the income-compensated demand, $D(P)$, we can calculate a monetary measure of the gain from skiing using the area under the compensated demand curve,

$$\phi(q) = \int_0^q D^{-1}(\tilde{q})\,d\tilde{q}.$$

This gain is the most that an individual would be willing to pay for the opportunity to ski q runs.

Individual i chooses to ski if the fixed cost, c_i, plus the explicit cost, $PD(P)$, is less than the gain from skiing, $\phi[D(P)]$. Our analysis allows the cost, c_i, to differ across persons and over time. On a given day, the cumulative distribution of the c_i's is described by F_s, so that the fraction of people who choose to ski is $F_s\{\phi[D(P)] - PD(P)\}$. Since $\phi[D(P)] - PD(P)$ is decreasing in P, the number of skiers falls if P rises. The shift parameter s represents changes over time in the distribution of the c_i. For example, during weekends and vacation periods, the costs of going skiing for the typical person are relatively low, so the number who ski is relatively high. Overall, we can write the number of persons N who choose to ski as the function,

$$N = N(P, s), \tag{1}$$

where $\partial N/\partial P < 0$.

By specifying that ski areas are competitive, we mean that each is small enough that its actions have a negligible impact on aggregate quantities. In this model (though not in those that follow), competitive behavior implies that firms take prices as given. Equilibrium requires that the total capacity of rides, Jx, equal the total number demanded, qN—that is,

$$Jx = D(P) \cdot N(P, s). \tag{2}$$

For a given value of Jx, this condition determines the equilibrium price per ride P. As one would expect, the price P falls with an increase in total capacity, Jx, and rises with an increase in the level of demand such as that generated from a downward shift in the fixed costs, c_i.

Over the longer term the model also determines the size of the industry, Jx. This scale depends on the cost of building new capacity (either more firms J or more rides per firm x) and on the distribution of returns, as determined by equation (2) and the distribution function of the shift parameter s.

B. *Equilibrium with Lift Tickets*

We now show how the equilibrium described above can be implemented using an entry fee (i.e., an all-day lift ticket) and a price per ride set equal to zero. Let π_j denote the price of a lift ticket at area j, and let n_j be the number of skiers who ski there. Given the total capacity x, the maximum number of rides per skier will be $q_j = x/n_j$. In equilibrium each person will desire a greater number of rides than x/n_j at the zero marginal cost implied by lift-ticket pricing. Hence, there is no problem in getting the customers to accept the quantity of rides available. In fact, people will queue up to receive the rides.

Each individual cares about the outlay on skiing, $c_i + \pi_j$, and the number of rides available, $q_j = x/n_j$. We assume that people do not care directly about the time spent waiting in lift lines, or about how the rides are distributed throughout the day. They would prefer shorter lift lines because they would

prefer more rides; but given a fixed number of rides, they are indifferent between spending time outdoors in line or indoors in the lodge. The only function of the queue is to allocate the fixed number of rides x equally among the n_j skiers.

As noted at the beginning of our paper, this extreme assumption about the welfare cost of time spent in lift lines is a useful expository device because it shows that queues may play a role that has nothing to do with the usual arguments about the cost of time. We do not take the welfare implications of this assumption literally. In Section I, Part C, we discuss departures from this assumption.

Suppose that individual i considers the choice between areas j and k, which offer the respective quantity of rides, $q_j = x/n_j$ and $q_k = x/n_k$. Since the cost c_i of going skiing is the same for each area, the individual will be indifferent between areas j and k if the gain from skiing minus the cost of the lift ticket is the same. Therefore, the equilibrium condition for people to be indifferent between areas is

$$\phi(q_j) - \pi_j = \phi(q_k) - \pi_k. \tag{3}$$

A ski area that is small relative to the total market can choose its lift-ticket price, π_j, but the number of skiers, n_j, adjusts to keep the net surplus, $\phi(x/n_j) - \pi_j$, equal to that offered by other areas. Competitive behavior means that the area takes as given a reservation value for the net surplus, not the price of the lift ticket. This given level of the net surplus implies a downward-sloping number of customers, n_j, as a function of π_j. The nature of the relation between n_j and π_j can be determined by implicit differentiation of the terms on the left side of equation (3). Using $\phi'(x/n_j) = D^{-1}(x/n_j)$, the result can be expressed in terms of the elasticity,

$$\frac{dn_j}{d\pi_j} \cdot \frac{\pi_j}{n_j} = -\frac{\pi_j}{D^{-1}(x/n_j) \cdot (x/n_j)}. \tag{4}$$

Since an area's costs are fixed and do not depend on n_j, each area seeks to maximize $\pi_j n_j$, taking as given the relation between the ticket price and the number of skiers implied by equation (4). As usual, maximization of revenue requires that the elasticity of n_j with respect to π_j equal -1, so that in equilibrium

$$\frac{\pi_j}{q_j} = D^{-1}(q_j). \tag{5}$$

The left side of equation (5) is the amount paid per ride under lift-ticket pricing, which we define as the effective price per ride, $\hat{P}_j$:

$$\hat{P}_j = \pi_j / q_j. \tag{6}$$

From equation (5) it follows that

$$q_j = D(\hat{P}_j). \tag{7}$$

Each person at area j ends up with the quantity of rides q_j that corresponds to the effective price per ride $\hat{P}_j$. Although people wait in line and face an explicit marginal cost for rides of zero, the results are as if each skier gets the quantity of rides that he or she would demand at an explicit market price per ride $\hat{P}_j$. These q_j rides have simply been combined into a package deal with a total cost of $\pi_j = \hat{P}_j q_j$.

Given the reservation value of net surplus, each area chooses its price, π_j, in accordance with equations (3) and (5) (taking account of the condition $q_j = x/n_j$). Since the areas have the same capacity x and are otherwise identical, they end up with the same values for the lift-ticket price, $\pi_j = \pi$, the number of customers, $n_j = N/J$, and the effective price per ride, $\hat{P}_j = \hat{P}$.

To complete the description of the equilibrium, it remains to determine the value of the common lift-ticket price, π, or equivalently of the effective price per ride $\hat{P}$. We can analyze individuals' decisions to incur the fixed cost to go skiing just as in the first model, except that the effective price per ride $\hat{P}$ now replaces the explicit price P. (Recall from equation (7) that people end up with the quantity of rides that they would demand at the effective price $\hat{P}$.) The analogue to equation (2) is now

$$Jx = D(\hat{P}) \cdot N(\hat{P}, s). \tag{8}$$

Because this condition is the same as the one that determined the price per ride P in the ride-ticket equilibrium, the effective price $\hat{P}$ takes on the same value. Finally, equations (6) and (7) imply that the common lift-ticket price is determined by the effective price per ride.

$$(9) \qquad \pi = \hat{P}q = \hat{P}\cdot D(\hat{P}).$$

Since the equilibrium with lift-ticket pricing yields an effective price per ride $\hat{P}$ equal to the explicit price per ride P in the first equilibrium, skiers receive the same number of rides at the same cost in each case. The same people end up participating, and each ski area receives the same revenue.

The equality of π_j and n_j across areas arises here because all ski areas and individuals are identical. Section III shows that π_j and n_j can vary across areas if there are differences in skiers' preferences or in the characteristics of ski areas. What generalizes is the result that lift-ticket pricing can replicate the quantities and marginal valuations generated under ride-ticket pricing. The lift-ticket equilibrium bundles the number of rides per person from the ride-ticket equilibrium into a single package that is sold at a price equal to the number of rides times the price per ride.

In the longer-run context, where the capacity Jx is variable, suppliers have the same incentives to invest under the two systems of pricing because the revenue generated by an additional ride corresponds in each case to the skiers' marginal valuation of rides, $D^{-1}(q)$. Thus—given our assumption that people care about the number of rides but not directly about the time spent in line —there are no inefficiencies implied by the existence of queues, which reflect the explicit marginal cost of zero for rides. Allocative decisions are still based on the proper shadow price, $\hat{P} = D^{-1}(q)$.

Although the lift-ticket equilibrium is only a repackaged form of the original competitive equilibrium, the superficial appearances are strikingly different. The lift-ticket solution features quantity rationing by means of queues, as well as ticket prices that seem to be set by firms with market power. Although individual ski areas have no true market power, the demand for lift tickets at each area is the downward-sloping curve $n_j(\pi_j)$ characterized by equation (4), and each area maximizes revenue subject to this curve.

C. *Ride Tickets vs. Lift Tickets*

Given the assumptions so far, there is no basis for predicting which of the two forms of pricing will prevail. They lead to identical allocations and effective prices. Ski areas charging on a per ride basis could coexist with others charging on a lift-ticket basis. One can readily verify that an area could also use a combination of a lift ticket (i.e., an entry fee) and a charge per ride.

The description of the world implicit in this model misses important features of reality. For some aspects, such as the determination of the price per ride P or $\hat{P}$, these features may not be important. However, in the choice between two otherwise equivalent pricing schemes, these features may be decisive. The most obvious elements neglected so far are: (*i*) the costs that must be incurred by an area to enforce contracts—for example, to avoid the theft of rides; (*ii*) the differences in rides—they are heterogeneous goods indexed by the time of day and by contingencies such as breakdowns of equipment and arrivals of skiers; and (*iii*) the time spent waiting in line, which is likely to have a positive opportunity cost.

We can conjecture what the inclusion of these features would imply. Given the allocation of rides common to the two kinds of equilibria (and to any mixture of these two), the form of pricing that minimizes the neglected costs will be selected. Ride-ticket pricing will generally have higher monitoring and set-up costs than lift-ticket pricing. Since it would be extremely expensive to set up a complete set of markets in time and contingency specific rides, and to enforce contracts written in this form, some amount of queuing would be expected even under pure ride pricing. On the other hand, lift-ticket pricing imposes costs in terms of time. Relative to a system with an extensive system of reservations, each individual must spend more time at a ski area to achieve a given

allocation of rides. However, if the typical skier's fixed cost, c, for getting to the ski area is large, and if waiting in line is preferred to spending time on other available activities (aside from skiing), then this last element would be relatively unimportant.

Queues also have the advantage that they permit an automatic form of *ex post* settling up to operate. Hence, transactions can take place before all the relevant contingencies have been realized, without the need for any *ex post* payments or recontracting. Consider the operation of ride-ticket pricing under the plausible assumption that there is uncertainty about the number of skiers who come to a ski area on a given day. To avoid the costs of repeated purchases of tickets throughout the day, skiers would presumably want to purchase all of their ride tickets for the day when they arrive. But if individuals arrive at a ski area and purchase tickets sequentially, the price for ride tickets offered to the first purchasers of the day will inevitably turn out to be incorrect *ex post*. For example, if more skiers than expected appear, too few ride tickets will remain for the late arrivals. If the ski area increases the explicit price per ride for late arrivals, early purchasers will have bought at a price that is too low; the marginal value of the last ticket held by an early purchaser will be less than what it could be sold for. Full efficiency would require trades between the early and late arrivals. Under lift-ticket pricing, the price per ride adjusts automatically. When more people show up—that is, n_j is larger —the effective price per ride, $\hat{P}_j = \pi_j n_j / x$, increases even if π_j is held constant. Section I, Part D, shows that in some cases this automatic price change is of exactly the right size.

As far as we know, ski areas use only the lift-ticket form of pricing. Walter Oi (1971) describes how Disneyland once followed a combination form of pricing with an entry fee and a charge per ride. In contrast to the explanation offered here, Oi interprets this scheme as evidence of market power. Disneyland has since shifted to a pure entry fee. We take these observations as evidence that the costs of allocating rides using ride tickets are higher than those using entry fees. Presumably, the cost of implementing reservations and collecting ride tickets outweighs the value of the savings in the time required to acquire a given number of rides. This outcome is likely if the lump-sum costs of participating are large, and if time spent at a ski area or amusement park is valued for its own sake.

D. *Shifts in Demand*

The foregoing arguments demonstrate that there may be little or no deadweight loss associated with the use of lift-ticket pricing, rather than ride-ticket pricing. But the results do not yet explain why ticket prices would be "sticky." Over the course of a season, variations in the shift parameter s—such as those reflecting weekends and vacation periods—cause predictable changes in demand. Lift lines vary markedly, as do prices for accommodations, but lift-ticket prices apparently change relatively little. The main variations seem to be discounts during periods of very low demand, such as nonvacation weekdays or the final days of the ski season.

As one would expect, equation (8) implies that the effective price per ride $\hat{P}$ varies in the same direction as the level of demand, with the sensitivity depending inversely on the magnitude of the price elasticity of the overall demand for rides (that is, of $D(\hat{P})\cdot N(\hat{P}, s)$). Thus, the effective price per ride is high when the level of demand is high, and vice versa. However, the price π for a lift ticket does not necessarily vary in the same direction as the level of demand. From equation (9), the effective price per ride is $\hat{P} = \pi/q = \pi n/x$. Even with π (and x) fixed, the extra crowding associated with the increase in n (which equals N/J) itself generates a higher effective price per ride. The lift-ticket price π increases when $\hat{P}$ increases only if the associated fall in rides per person, $x/n = D(\hat{P})$, is less than equiproportional. Using equation (9), the effect of a change in $\hat{P}$ on π is

$$(10) \quad d\pi/d\hat{P} = D(\hat{P})(1+\eta_{D,\hat{P}}) \gtreqless 0,$$

where $\eta_{D,\hat{P}} < 0$ is the elasticity of rides de-

manded per person with respect to the effective price per ride. If this elasticity is less than 1 in absolute value, π rises along with $\hat{P}$ and, hence, with the level of demand. But if the elasticity is greater than 1 in absolute value, π falls when $\hat{P}$ increases. Finally, if the elasticity is close to -1, π shows little sensitivity to fluctuations in demand.[2] In this case competitive forces are consistent with nearly constant lift-ticket prices, even though the times of peak demand exhibit lines that are much longer than those during nonpeak times.

This result suggests an additional advantage to lift-ticket pricing. If the elasticity of demand for rides per person is close to -1, it is unnecessary to incur the "menu costs" of changing the stated price at a ski area in response to changes in demand. The effective price per ride changes in nearly the right way if the price of lift tickets is held constant. As suggested in Section I, Part C, this may be important even over the course of a day. If ticket sales take place sequentially as customers arrive, the cost of changing the ticket price as information on the size of demand accumulates includes not just a menu cost of changing signs, but also the costs of recontracting with previous purchasers.

If demand falls to very low levels, the condition $|\eta_{D,\hat{P}}|<1$ almost surely applies. Since the consumption of a ski run requires a minimum amount of a skier's time, the demand curve for rides as a function of the price has a finite intercept equal to the maximum number of ski runs that can be taken in a given day. As the effective price of a ski run approaches zero, the elasticity of demand must also approach zero. In this region equation (10) implies $d\pi/d\hat{P}>0$—that is, the model predicts discounts on lift tickets during the times of greatest slack. However, the model also suggests the possibility of a substantial interval—such as the comparison between a normal weekend and the peaks during vacation periods—where lift-ticket prices would show little or no variation with demand.

The same mechanism may explain why the explicit prices for goods such as airline tickets and restaurants often do not vary between peak and off-peak periods. At busy times the effective amount of service diminishes because planes and restaurants are more crowded. Thus, the price per effective unit of service rises automatically if the explicit price is held fixed. Under such circumstances, the results with fixed explicit prices may roughly replicate the equilibrium where the price per effective unit of service fluctuates and where customers are free to choose how much service to purchase. In these examples the package-deal effect operates with quantity constraints that are implicit rather than explicit, and the results do not depend on queues per se.

Constant lift-ticket prices work exactly only if the elasticity of the demand for rides per person equals -1. But if the menu costs or the costs of recontracting due to sequential service are large enough to play a decisive role in the choice of the pricing format, a two-part pricing scheme with an entry fee and a price per ride can be implemented to avoid price changes even when the elasticity differs from -1. This consideration does not appear to be relevant for ski areas, where charges per ride do not seem to be used, but may be a factor in the choice of such a scheme by some amusement parks.

Consider an amusement park with capacity x, which charges an entry fee π and a price per ride r. If n people visit the park and the value of r is small (so that π is positive in equilibrium), the quantity of rides per person is $q=x/n$. As before, each park takes as given the net surplus attained by participants, as given now by $\phi(x/n_j)-rx/n_j-\pi$. Setting the total differential of the net surplus to zero and equating the elasticity of n with respect to π to -1 leads to

$$r+\frac{\pi}{q}=D^{-1}(q), \tag{11}$$

which extends equation (5). The effective

[2] The same mechanism works for shifts in supply, Jx, although these seem less important in the short run in the present context.

price per ride is now[3]

$$\hat{P} = r + \frac{\pi}{q} = r + \frac{\pi}{D(\hat{P})}. \tag{12}$$

Solving for π gives

$$\pi = (\hat{P} - r)D(\hat{P}). \tag{13}$$

As before (equation (8)), the equation of total supply to total demand determines $\hat{P}$. Then the reaction of π to changes in $\hat{P}$ follows from equation (13) as

$$d\pi/d\hat{P} = D(\hat{P})\left[1 + ((\hat{P} - r)/\hat{P})\eta_{D,\hat{P}}\right]. \tag{14}$$

Consider small fluctuations in demand or supply that induce fluctuations in the effective price $\hat{P}$ around some level $\hat{P}_0$. Let $\eta_{D,\hat{P}}$ be the elasticity of the demand for rides with respect to the effective price. For a given value of $\eta_{D,\hat{P}}$, the price per ride r can be chosen so that $((\hat{P} - r)/\hat{P})\eta_{D,\hat{P}}$ is equal to -1 when evaluated at $\hat{P} = \hat{P}_0$.[4] Then $d\pi/d\hat{P}$ will be zero when evaluated at $\hat{P}_0$ and it will be small in a neighborhood of $\hat{P}_0$. For small fluctuations in demand or supply, an equilibrium with constant prices r and π will be approximately equivalent to the conventional equilibrium with no entry fee and a fluctuating price P per ride.

II. Elaborations of the Model

In Section I, ski areas were identical and individuals differed only in terms of the fixed cost c_i; conditional on participation, they too were identical. In Section II, we illustrate how the previous results change when there are differences in characteristics of ski areas and in individuals' preferences for ski runs. Differences among ski areas lead to results that complement those above about sticky prices. The conditions that cause lift-ticket prices to be invariant with demand also cause these prices to be the same at areas with different characteristics. By considering two distinct kinds of differences across areas, we are also able to illustrate more clearly the separate roles of the homogeneity effect and the package-deal effect.

A. *Differences in Qualities of Ski Areas*

The lift-ticket equilibrium derived above does not require separate tickets at each area. Suppose that the ski operators set a single entry fee, which equals the common lift-ticket price. Then skiers will sort themselves so that each area has the same number of skiers. This sorting sets the average return to attendance at each area to a single value, and simultaneously sets the marginal value of an additional ski run at each area to a different common value. Since the areas are identical, there is no conflict between equating marginal and average quantities. In terms of the analogy with the classical two-roads problem, if the roads are identical and individuals are free to choose between them, the number of vehicles on each will be the same and there will be no efficiency loss from misallocation of cars between the roads. Identical roads or ski areas is a special case, but this section shows that this result generalizes to allow for at least one kind of difference across areas.

For simplicity we now suppress the participation decision and consider a pool of N identical skiers who have decided to go skiing. Skiers choose among J areas, which now have different effective capacities x_j. These differences could reflect variations in lift capacities or in lengths of ski runs.

As in Section I, we can derive the equilibrium for this extended model under lift-ticket pricing. All areas charge the same lift-ticket price, but the number of skiers n_j varies one for one with the capacity x_j. Each skier receives the same amount of skiing, x_j/n_j, and this amount coincides with the

[3] If $\hat{P} > r$, which we assume, the quantity demanded at the explicit price r exceeds the amount available, $q = x/n$. Therefore, although the explicit price is now positive, the demanders still queue up for the available rides. These queues typically applied at Disneyland even when ride tickets were used.

[4] However, the value of r would be negative if $\eta_{D,\hat{P}}$ were less than 1 in absolute value.

quantity that each would receive if the operators charged an explicit price per unit of skiing. Moreover, the results would be the same if the operators levied a single entry fee for skiing and allowed the skiers to allocate themselves among the areas.

To reconcile this result with intuition about congestion costs, consider the analogues to average and marginal costs. A single lift ticket combined with free movement among areas means that the surplus per person, $\phi(x_j/n_j)$, and hence the amount of skiing per person, x_j/n_j, are the same at all areas. On the other hand, a social planner who allocates skiers across areas would seek to maximize the total gain from skiing, $\Sigma n_j \phi(x_j/n_j)$, subject to $\Sigma n_j = N$. The first-order condition for this problem is that the expression, $\phi(x_j/n_j)-(x_j/n_j)\phi'(x_j/n_j)$, be the same at all areas. This condition also implies that the amount of skiing per person is the same at all areas. Hence, the allocation of skiers coincides with the one chosen privately. Since $n\phi(x/n)$ is homogeneous of degree 1 in x and n, the finding is a special case of the result noted at the beginning of our paper; if $g(x,n)$ is homogeneous of degree 1, then equating the average product, g/n, leads to the same answer as equating the marginal product, $\partial g/\partial n$.

This argument applies also within a given ski area or amusement park. Ski areas may not have to charge different prices for runs of different lengths or qualities, and amusement parks may not have to charge different prices for rides with different durations or levels of excitement. If a ride on a roller coaster is x times more satisfying than one on a bumper car, lines at the two activities will adjust so that each person can consume x times as many rides per day on a bumper car as compared to a roller coaster.

B. *Differences in Transportation Costs*

Now suppose that ski areas can differ by their costs of access. To simplify matters, assume again that the areas have the same capacity x. Let b_j denote the cost for any skier to travel to area j; for example, b_j could depend on the distance of the area from a major urban center. As before, π_j is the lift-ticket price, $q_j = x/n_j$ is the amount of skiing received by each person, and $\hat{P}_j = \pi_j/q_j$ is the effective price per ski run. By extension of equation (3), an individual is indifferent between areas j and k, if

$$(15)\quad \phi(q_j)-\pi_j-b_j=\phi(q_k)-\pi_k-b_k.$$

As before, a change in the lift-ticket price, π_j, causes the number of skiers, n_j, to adjust so that the net surplus on the left side of equation (15) remains constant. Hence, as in equation (5), revenue maximization by the firm implies $\pi_j = q_j D^{-1}(q_j)$. Inserting this result into equation (15) gives

$$(16)\quad \phi(q_j)-D^{-1}(q_j)q_j-b_j = \phi(q_k)-D^{-1}(q_k)q_k-b_k.$$

The term $\phi(q_j)-D^{-1}(q_j)q_j$, in equation (16) is the standard measure of consumer surplus—that is, the maximum amount that a consumer would pay for the privilege of buying q_j rides at price $D^{-1}(q_j)$. The equation says that this measure of consumer surplus at the two areas must differ by the difference in the transportation costs, $b_j - b_k$. For example, suppose that area j is closer than area k, so that $b_j < b_k$. Then, since consumer surplus is increasing in q, equation (16) implies $q_j < q_k$, and hence $\hat{P}_j = D^{-1}(q_j) > \hat{P}_k = D^{-1}(q_k)$. Thus, closer areas have higher effective prices per ski run and are more crowded in the sense of offering fewer rides per person.

Given the determination of $\hat{P}_j$, the solution for π_j follows from $\pi_j = D^{-1}(q_j)q_j = \hat{P}_j D(\hat{P}_j)$. Thus, the relation between lift-ticket prices, π_j, and the cost of access, b_j, depends again on the elasticity of the demand for rides per person. A lower value of b_j implies a higher value of $\hat{P}_j$ and hence a higher value of π_j if the elasticity of $D(\hat{P})$ with respect to $\hat{P}$ is less than one in magnitude (in the relevant range of demand). But a low b_j implies a low π_j (along with a high $\hat{P}_j$) if the elasticity exceeds one in magnitude. Finally, if the elasticity equals -1 in some range, then lift-ticket prices do not vary in this range across areas that differ in their costs of access.

Except when the elasticity is equal to -1, different areas must charge different amounts for lift tickets. A single entry fee for skiing with free choice among areas will not achieve the social optimum because the homogeneity effect does not operate. To see why, note that the total output from area j, net of transportation cost, is $h(b_j, n_j) = n_j\phi(x/n_j) - b_j n_j$. This function, $h(b_j, n_j)$, is not homogeneous of degree one in b_j and n_j, and cannot be written in the form, $h(n_j, b_j) = b_j f(n_j)$, for some homogeneous function $f(n_j)$.

Section II, Part C, shows that differences in individual preferences cause lift-ticket prices to differ among areas. But to the extent that these differences are small, the present results permit a kind of cross-sectional check on the explanation proposed above for the stickiness of lift-ticket prices. Many explanations can be offered for price stickiness over time, but it is harder to explain cross-sectional stickiness. If the demand curve for ski runs per person is close to unit elastic, then there should be less variation in lift-ticket prices than in the number of skiers or the length of lift lines, both in comparisons over time and among areas at a point in time. In both dimensions, it will appear that quantities respond more than prices.

C. *Differences in Preferences*

Suppose now that the demand curves for lift rides, $D_i(P)$, differ across individuals. These differences could reflect variations in preferences or incomes. At a given effective price, $\hat{P}$, the quantity of rides demanded per person differs from one person to another. All of the previous equilibria with lift tickets used a queuing mechanism to deliver the same number of rides to all skiers. Since this mechanism does not discriminate among people with different preferences, it cannot allocate different numbers of rides per day, $D_i(\hat{P})$, to them. To achieve an allocation that does discriminate, different areas (or different classes of tickets at a single area) will have to cater to different types of individuals.

To illustrate the results, suppose that there are two types of customers. Avid skiers have the demand $q^A = D^A(P)$, while less avid skiers have the demand $q^B = D^B(P)$, where $q^A > q^B$ for any value of P. The J ski areas will end up dividing themselves into two types: a number J^A that caters to type A customers, and a number $J^B = J - J^A$ that serves the type B customers. Given the numbers J^A and J^B, the determination of effective prices per ski run, $\hat{P}^A$ and $\hat{P}^B$, and lift-ticket prices, π^A and π^B, proceeds as before. In particular, assuming that A-type skiers go to A-type areas, and similarly for B types, the conditions are

$$N^A(\hat{P}^A, s)\cdot D(\hat{P}^A) = J^A x$$

$$\pi^A = \hat{P}^A \cdot D(\hat{P}^A)$$

and analogously for the B's.

Ski areas can choose between proclaiming themselves as type A, with revenue $\hat{P}^A x$, or type B, with revenue $\hat{P}^B x$. Hence, the numbers J^A and J^B adjust in an equilibrium to attain $\hat{P}^A = \hat{P}^B = \hat{P}$.[5] In that case, we find

$$(17) \qquad q^A = D^A(\hat{P}) > q^B = D^B(\hat{P})$$

and

$$(18) \qquad \pi^A = \hat{P}\cdot D^A(\hat{P}) > \pi^B = \hat{P}\cdot D^B(\hat{P}).$$

That is, more avid skiers receive more rides and pay higher lift-ticket prices.[6]

Recall that $D^A(\hat{P}) = x/n^A$ and $D^B(\hat{P}) = x/n^B$, where n^A and n^B are the number of skiers at each type of area. Hence, equation (18) implies

$$(19) \qquad \frac{n^A}{n^B} = \frac{\pi^B}{\pi^A}.$$

In other words, in deciding whether to charge

[5] We neglect integer constraints on the solution. If the number of areas (that serve a given locality) is large, then this problem would be unimportant. Also, differences in capacities, x_j, make this issue less serious.

[6] Note that type A skiers prefer q^A rides at price π^A to q^B rides at price π^B (since $q^A = D^A(\hat{P})$ is the quantity demanded at the effective price $\hat{P}$). Similarly, type B skiers prefer q^B rides at price π^B to q^A rides at price π^A. Therefore the separating equilibrium that we propose is viable.

π^A or π^B—that is, whether to be a type A or type B ski area—each area faces a demand in terms of numbers of skiers, n, that has an elasticity of precisely -1 with respect to the lift-ticket price. Correspondingly, the area's revenue, πn, is invariant with the choice among the π's. The areas are indifferent between charging a high lift-ticket price and catering with short lines to the skiers who demand lots of rides per person, or charging a low price and servicing with long lines those who demand few rides. Note also that, as an area changes its lift-ticket price, it also changes the entire class of skiers that choose to patronize it. That is, a shift from π^A to π^B means that the n^A previous customers all leave, while n^B new customers show up.

The results generalize to multiple-skier types, which lead to multiple values of q^k and π^k, but still a single value of $\hat{P}$.[7] A ski area's revenues must still be invariant to its choice of type—that is, of π^k. Therefore, over the set of values for π^k that prevail in equilibrium, it again follows that each area faces a demand in terms of number of skiers, n^k, that has an elasticity of -1 with respect to π^k. This equiproportional change in the number of skiers in response to a change in the lift-ticket price does not depend on the elasticity of the aggregate demand for lift rides or of individuals' demands for rides per person. The result obtains whenever a range of lift-ticket prices π^k prevails in equilibrium.

Except for the restriction to a finite number of individual types, and hence a finite number of observed prices π^k, the lift-ticket equilibrium in the presence of different tastes resembles the equilibrium with differentiated products and hedonic prices as described in Sherwin Rosen (1974). Each type of ski area offers a different type of skiing experience, indexed by q^k, the number of ski runs available per skier. With identical competitive producers, profit is invariant to the type of good offered, and the price function $\pi(q)$ traces out the structure of the demand side of the market. In Rosen's model, producers with no market power choose the type of good offered and the price charged from the locus described by $\pi(q)$. Here, the departure from the standard model of competitive price taking is even sharper. Firms simply choose a price π; quality—that is, the number of skiers—adjusts endogenously. It is interesting to note that, until recently, the Metro in Paris used a similar scheme to sort people by tastes. Purchasers of first-class tickets rode in separate cars, which were not physically different from second-class cars, except that the first-class cars were less crowded (in equilibrium).[8] Roughly speaking, individuals with a stronger preference for ski rides or with a greater distaste for congestion are willing to pay more for the opportunity to pay more.

III. Applications to Other Markets

In this section, we apply the paradigm of ski-lift pricing to two classic problems of congestion, the two-roads problem noted above and an open-access fishing problem. We conclude with an application to employment contracts with profit sharing. Our objective is to illustrate the applicability of the approach to a variety of problems, and to use some well-known examples to clarify the distinction between the package-deal effect and the homogeneity effect.

A. *Classical Congestion*

Suppose that there are two roads that connect a pair of cities. Let $v(x,n)$ describe the speed of cars traveling on each road as a function of the capacity x of each road and

[7]However, the integer problem mentioned in fn. 5 becomes more serious when there are multiple types. If the number of types is much greater than the number of ski areas, then each area will have to cater to a range of skier types. In this case the use of lift tickets will involve an additional welfare loss relative to an equilibrium with ride tickets.

[8]We are told that the abolition of this vestige of the class system was one of the promises made in the presidential campaign of Francois Mitterand. After his election, a compromise was reached whereby this system was not allowed to operate during the morning and evening periods of peak demand (where it presumably would be most useful), but remained in effect during the middle of the day.

the number of cars n. Thus, $h(x,n)=nv(x,n)$ is the rate of flow of cars along each road. As Knight (1924) pointed out, if no toll is charged for the use of either road, individuals will sort themselves so that the average output, $h(x,n)/n$, is the same on the two roads; that is, the speeds and travel times will be the same. However, social optimality requires that the aggregate travel time summed over all individuals be minimized, which is equivalent to maximizing the total flow, $h(x_1,n_1)+h(x_2,n_2)$. As noted at the beginning of our paper, the private and social choices coincide if h is homogeneous of degree 1, or if it can be written in the form, $h(x,n)=xf(n)$, where f is homogeneous of some degree. The first specification implies that speed depends on the relationship of capacity to the number of cars, but is homogeneous of degree 0. Thus, as seems reasonable, doubling the capacity of the road and the number of vehicles leaves the speed unchanged.

The suboptimality studied by Knight relied on the assumption that one of the roads had a capacity that was so large that the speed was independent of the number of cars, $v_1(n)=a$. The second road was assumed to be subject to congestion; for example, $v_2(x,n)=x_2f(n)$ for some decreasing function f. The free-access equilibrium is then suboptimal, but no justification was given for the different functional dependencies on capacity.

Consider now the case of J fishing ponds, with n_j fishermen at pond j. For the moment, we treat the total number of fishermen, $N=\Sigma n_j$, as fixed. We assume that the output of fish at pond j takes the form

$$y_j=x_j(n_j)^{\alpha}, \tag{20}$$

where x_j is the intrinsic quality of the pond and $0\leq\alpha<1$. The condition $\alpha=0$ corresponds to the case of a ski lift with fixed capacity, x_j. The case $\alpha>0$ means that an additional fisherman raises the total catch, but if $\alpha<1$, the marginal and average product diminish with n_j. Thus the pond is subject to congestion; adding an additional fisherman reduces the catch of the previous fishermen. Now suppose that each pond has the same value of α—that is, although the x_j's can vary, crowding sets in at the same proportionate rate at each pond.

If there are no admission fees, each fisherman goes to the pond that promises the highest average product, so that in equilibrium, the average products, $x_jn_j^{\alpha-1}$, are equal at each pond. For fixed N, a social planner would seek (in this static problem) to maximize the total current output of fish, $Y=\Sigma y_j$. This maximization requires the marginal product, $\alpha x_jn_j^{\alpha-1}$, to be equal at each pond. But this condition generates the same number, n_j, as the private choices. In other words, despite the congestion problem, the decentralized solution with no explicit prices achieves a Pareto optimal allocation of fishermen.

The result depends on the assumption that crowding sets in at the same proportionate rate at each pond, as implied by the form of equation (20). To see this, relax the assumption that N is fixed, and assume instead that fishermen have a distribution of costs for going fishing, c_i. This means that each person has available an alternative activity—such as staying home—where the output does not involve the same sort of crowding as prevails at each fishing pond. In the decentralized solution a person chooses to fish if $c_i<$ APL, where APL is the common value of the average product of labor. (This condition applies to commercial fishing and assumes no utility from fishing, per se.) On the other hand, the social planner would assign a person to fishing if $c_i<$ MPL, where MPL is the common value of the marginal product. Since MPL < APL, we get the standard result that too many people choose to fish under the decentralized solution. However, to attain a Pareto optimum, it is necessary to charge only a single price π to fish—that is, a fishing license.[9] It is unnecessary to have

[9]The price is $\pi=\text{APL}-\text{MPL}=(1-\alpha)x_j(n_j)^{\alpha-1}$, where j is any of the ponds, and the total number of fishermen N is the number for whom $c_i+\pi<\text{APL}=x_j(n_j)^{\alpha-1}$ (or $c_i<\text{MPL}=\alpha x_j(n_j)^{\alpha-1}$). If there is a downward shift in the c_i's, then N, and hence n_j, rise, which implies a decline in π. Hence, with shifts in labor supply, fishing licenses should be cheaper when the ponds are more crowded.

different prices at the various ponds, even though they differ by their intrinsic qualities, x_j. In equilibrium the better ponds are more crowded—but to exactly the extent required to attain the optimal allocation of fishermen.

The fishing problem has an analogue to the assignment of people to rooms for sessions at a professional meeting. Sessions differ by their intrinsic quality, x_j. However, as more people crowd in, it becomes more difficult to see or hear, so that the "quantity" received per person declines with n_j. For example, if crowding sets in at the same proportionate rate at each session, then equation (20) describes the total "output" of session j. If the total number of participants, N, is fixed, then the decentralized choices achieve a Pareto optimum without having explicit prices for each session. If the number N is variable—in particular, if people have access to some alternative activity that is not subject to congestion—then too many people attend sessions in a decentralized equilibrium. However, the attainment of a Pareto optimum requires only a single fee (a registration charge), and not prices that vary across sessions in accordance with their intrinsic qualities.

B. *Profit Sharing*

In the labor market a fully flexible wage rate corresponds to a flexible price per lift ride. The case of a lift ticket relates to alternative methods of labor compensation, such as profit-sharing schemes. From the standpoint of an individual worker, the firm's total profit looks like a ski operator's total capacity. In particular, the amount that each person gets (share of profits or number of lift rides) varies inversely with the number of other people who show up. (Profit per worker falls with more workers as long as the average product of labor exceeds the marginal product.) But, as in the ski example, competition among firms causes the parameters of the profit-sharing rule to adjust so as to reproduce the outcomes that would emerge under flexible wages. Further, under some conditions, it would be satisfactory to have fixed wages and fixed parameters for the profit-sharing formula.

Suppose that each of J identical, competitive firms has the production function,

$$Y = A \cdot F(n), \tag{21}$$

where Y is output, A is a technological shift parameter, and n is the number of workers. We assume that each worker has the same productivity and works a fixed number of hours per day. We deal initially with a standard setting where the real wage rate, w, is flexible. Given this wage, profit maximization for each firm entails

$$AF'(n) = w. \tag{22}$$

Equation (22) determines each firm's labor demand. Aggregate labor demand is the multiple J of the demand per firm.

The economy has a population of M potential workers who have a distribution of reservation wages, c_i; those with $c_i < w$ choose to work. Hence, the aggregate labor supply function is

$$N = N(w, \theta), \tag{23}$$

where $N \leq M$, $(\partial N/\partial w) \geq 0$, and θ represents factors (including wealth effects) that influence the position of the distribution of reservation wages.

The equation of aggregate labor supply to aggregate labor demand determines the market-clearing values of the wage rate, w^*, and employment, N^*. Then each firm's employment is $n^* = N^*/J$. We assume that variations over time in wages rates and employment reflect shifts in the technological parameter, A, or in the reservation-wage parameter, θ. (A shift in the parameter A, to the extent that it changes wealth, could imply a simultaneous shift in θ.)

As in the ski-lift example, the competitive equilibrium in the labor market can be supported by pricing mechanisms other than the obvious one of freely flexible wage rates per worker. For example, assume that the wage rate is fixed at some value $w' < w^*$. The wage w' parallels the explicit price per ride, r, in the ski-lift case. Therefore, $w' = 0$ corresponds to pure lift-ticket pricing, where $r = 0$.

Assume now that each worker's compensation consists of the wage w' plus a bonus B. We consider a profit-sharing scheme, similar to Martin Weitzman (1985), where the bonus to each worker is the fraction β of profit per worker; that is,

$$(24)\qquad B=\beta\left[\frac{AF(n)}{n}-w'\right],$$

where $0\leq\beta\leq 1$. Therefore, a worker's total compensation, Ω, is given by

$$(25)\quad \Omega=B+w'=\beta AF(n)/n+(1-\beta)w'.$$

Since the potential workers care only about Ω, and not on its division between B and w', each competitive firm takes as given the value of Ω that it must pay. Hence, for fixed w', equation (25) determines how the quantity of labor supplied to the firm, n, varies with the profit-sharing fraction, β. That is, each firm can call out a value of β (along with an arbitrary w'), and the number n adjusts to make the overall compensation, Ω, equal to its competitively determined value. This adjustment of n to a change in β parallels the response of the number of skiers to a shift in the lift-ticket price.

Setting the differential of Ω in equation (25) to zero leads to

$$(26)\quad \frac{dn}{d\beta}=\frac{F(n)-nw'/A}{\beta\left[\frac{F(n)}{n}-F'(n)\right]}>0.$$

The denominator is positive from the usual assumptions about the production function —that is, the average product, $AF(n)/n$, exceeds the marginal product, $AF'(n)$. The numerator is positive if, as we assume, w' is less than the average product, $AF(n)/n$.[10]

Each firm chooses β to maximize profit, as given by $(1-\beta)[AF(n)-nw']$, subject to the relation between n and β from equation (26). Setting to zero the derivative of profit with respect to β (taking account of the response of n from equation (26)) leads to the condition

$$(27)\qquad AF'(n)=\Omega.$$

That is, labor's marginal product equals the total compensation, Ω, not the explicit wage w'. Hence, with the substitution of Ω for w, the result parallels that with a flexible wage in equation (22). In particular, labor demand, n, depends on Ω exactly as it depended before on w.

Potential workers participate in the market if their reservation wage, c_i, is below the total compensation Ω. Therefore, the aggregate labor supply function is now

$$(28)\qquad N=N(\Omega,\theta),$$

which parallels the flexible-wage case in equation (23), except for the substitution of Ω for w.

Equations (27) and (28) imply that aggregate labor demand and supply depend on Ω exactly as they depended on w in the flexible-wage case. Therefore, the market-clearing value of total compensation, Ω^*, equals the market-clearing flexible wage rate, w^*. It follows that all allocations—including employment per firm, n^*, and labor-force participation, N^*—coincide with those in the flexible-wage case.

Note that, although the allocations are the same, the appearances are again quite different. Under the bonus arrangement, the explicit wage, $w'<w^*$, is rigid, but each worker also gets a share of the profits. In deciding whether to work, each person looks only at the total compensation, $\Omega=w'+B$, and neglects the negative effect of his participation on the profit distributed to the other workers (which occurs because the marginal product of labor is below the average product). This interaction parallels the negative effect of an additional skier's participation on the rides available for others. Nevertheless, as in our previous example for skiing, the profit-sharing scheme reproduces the results for employment and total compensation per worker that would arise under flexible wages.

It also follows that firms would eagerly hire more workers than are available at the

[10] The assumption $w'<w^*$ turns out to guarantee this condition at the equilibrium value of n.

going wage $w'(<w^*)$ and the prescribed terms for sharing profits. (This result parallels the eagerness of skiers to queue up for lift rides.) But more workers than n^* do not present themselves because the total compensation, $\Omega = w' + B$, would then fall below the value w^*, which is the reservation wage of the marginal worker (when total employment is $N^* = Jn^*$). As with flexible wages, employment is determined so that labor's marginal product equals the competitive wage w^*. In other words, profits are maximized subject to the constraint that firms pay each worker a total compensation that equals the marginal worker's reservation wage. Thus, the outcomes are Pareto optimal despite rigid wages and the apparent common-property problem associated with the sharing of profits. Even though the marginal cost to the firm of an additional worker is less under the profit-sharing arrangement, the equilibrium level of employment is the same as that with flexible wages. Correspondingly, the firms in each regime face the appropriate shadow price of labor (w^*), and thereby make correct decisions with respect to investment in capital, entry and exit, etc.

All of the results so far parallel those from Section II. In particular, they depend only on the package-deal effect. The package offered here—in this case by a buyer of labor services—is w' of wage dollars plus $B = w^* - w'$ of bonus dollars per unit of labor.

The discussion of amusement parks noted that, for any given elasticity of demand, there would exist a constant entry fee and a constant price per ride such that small shocks generated outcomes that approximated those from a flexible price equilibrium. A similar local result holds here. There will exist fixed values of w' and β such that disturbances generate outcomes that approximate those supported by flexible prices.

There is also an interesting special case in which the parameters, β and w', can be constant in the face of global shocks to supply or demand. (This parallels the ski-lift example in which the lift-ticket price does not vary with shocks if the elasticity of demand for rides per person equals -1.) By substituting for Ω from equation (25), the equilibrium condition from equation (27) is

$$(29)\quad AF'(n) = \beta AF(n)/n + (1-\beta)w'.$$

A change in the reservation-wage parameter, θ, leads to a change in n, which leads generally to a shift in β for a given w'. Total differentiation of equation (29) with respect to n and β (for fixed A and w') leads to

$$d\beta/dn = H\left[\beta F(n) - \beta nF'(n) + n^2F''(n)\right],$$

where H is a positive expression. This derivative will equal zero for all values of β and n if the production function has the Cobb-Douglas form, $F(n) = n^{\alpha}$, and if β is set equal to α.[11] It can be verified from equation (29) that, for this production function, the share parameter β also does not change with a shift in the technological parameter, A. Substitution of $F(n) = n^{\beta}$ into equation (29) shows that the level of total compensation is correct here only if $w' = 0$. Hence, if production is Cobb-Douglas with labor's share β, then the firms can pay workers a zero explicit wage, w', and a fraction β of the profits. Under this scheme, the values of β and w' do not have to change with variations in labor supply or proportional shifts to the production function in order to support the competitive allocations.

The present results do not imply that profit sharing is superior to other schemes that allow the bonus (and thereby total labor compensation) to move along with w^*. Also, the analysis does not suggest that a framework with fixed wages and a flexible bonus (related, say, to profits) would be superior to a setup with flexible wages. As was the case for ski areas, the choice of compensation scheme must be based on elements of reality that are excluded from this model.

[11] For $\beta \neq 0$, the general solution for $d\beta/dn = 0$ is $F(n) = c_1 n^{\beta} + c_2 n$, where c_1 and c_2 are arbitrary constants. However, β varies with shifts in the parameter A unless $c_2 = 0$.

REFERENCES

Barro, Robert J., "Long-Term Contracting, Sticky Prices, and Monetary Policy," *Journal of Monetary Economics*, July 1977, *3*, 305–316.

Hall, Robert E., "Employment Fluctuations and Wage Rigidity," *Brookings Papers on Economic Activity*, 1: 1980, 91–123.

Knight, Frank F., "Fallacies in the Interpretation of Social Cost," *Quarterly Journal of Economics*, August 1924, *38*, 582–606.

Oi, Walter Y., "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly," *Quarterly Journal of Economics*, February 1971, *85*, 77–96.

Rosen, Sherwin, "Hedonic Prices and Implicit Markets," *Journal of Political Economy*, January/February 1974, *82*, 34–55.

Weitzman, Martin L., "The Simple Macroeconomics of Profit Sharing," *American Economic Review*, December 1985, *75*, 937–53.

Strategic Behavior in Contests

By AVINASH DIXIT*

This paper considers the effect of precommitment in contests where the rivals expend effort to win a prize. With two asymmetric players, it is found that the favorite will commit effort at a higher level than that in a Nash equilibrium without commitment, and the underdog at a lower level. With many players, the absence of an odds-on favorite among rivals is sufficient to ensure overcommitment by any one player. Applications to sports, oligopoly, and rent seeking are discussed.

Many economic and social games are contests where the players expend effort to increase their probability of winning a given prize. Examples include: (*i*) inter-firm or international *R&D* rivalry for a profitable innovation; (*ii*) bribery to secure a lucrative license or contract from a government; and (*iii*) the Wimbledon final or the Superbowl. Economists have studied such games in many contexts. Glenn Loury, 1979, Partha Dasgupta and Joseph Stiglitz, 1980, and Tom Lee and Louis Wilde, 1980, have examined *R&D* rivalry. Gordon Tullock, 1980, has analyzed rent seeking. John Riley, 1979, and Barry Nalebuff and Riley, 1985, have dealt with wars of attrition. Ed Lazear and Sherwin Rosen, 1981, Jerry Green and Nancy Stokey, 1983, Bengt Holmstrom, 1982, Nalebuff and Stiglitz, 1983, and Sherwin Rosen, 1986, have looked at contests from the viewpoint of incentive design. Dale Mortensen, 1982, has studied a general model with different interpretations and applications.

In all of that literature, the focus is on the existence and the characterization of Nash equilibria. In this paper I examine a different aspect. Suppose one of the players is given the opportunity to make a strategic precommitment to the level of effort. This may be done directly by imposing an order of moves (Stackelberg leadership), or indirectly through some other variable that will influence the *ex post* choice of effort (such as capacity or tax-subsidy policy in industrial games), or merely by psyching oneself up (or down) in sporting contests. The question I address is: What governs whether a player will choose an effort level that is higher or lower than the Nash equilibrium level without precommitment?

I find that in a perfectly symmetric two-player game, there is no *local* incentive for any strategic manipulation of the effort level. Analysis of the case in which the players have asymmetric prospects requires special forms of the function governing the probabilities of winning. For the two specifications most commonly used in the literature, namely logit and probit, there is a remarkably simple answer. The player who is the favorite (in the sense of having a probability of winning greater than $1/2$ in the Nash equilibrium) has the incentive to overcommit effort, and the other has the opposite incentive. Section I sets up the model and establishes these results.

Section II introduces several contestants, of whom one is given the opportunity to precommit effort. For the logit specification, there is a simple sufficient condition for strategic overcommitment of effort by this player: there should be no odds-on favorite among the rest. Therefore it appears that overexertion is the likely case in most contests of several players in practice. Section III examines some applications and extensions of the basic model.

*Department of Economics, Princeton University, Princeton, NJ 08544. I thank Barry Nalebuff for valuable discussions, Carl Shapiro, Robert Solow, and the referee for helpful comments, and the National Science Foundation for research support under grant no. SES-8509536.

I. The Two-Person Case

Let K be the prize, x_1, x_2 the two players' effort levels in units commensurate with the prize, and $p(x_1, x_2)$ the probability that player 1 wins. Then the players' expected payoffs are

$$\pi_1 = Kp(x_1, x_2) - x_1 \tag{1}$$

and

$$\pi_2 = K[1 - p(x_1, x_2)] - x_2. \tag{2}$$

To ensure a positive but diminishing marginal effect of each player's effort on his own probability of winning, I assume

$$p_1 > 0, \quad p_{11} < 0, \quad p_2 < 0, \quad p_{22} > 0, \tag{3}$$

where subscripts denote partial derivatives and the arguments (x_1, x_2) are omitted for brevity.

The first-order conditions for Nash equilibrium are

$$\partial\pi_1/\partial x_1 \equiv Kp_1(x_1, x_2) - 1 = 0 \tag{4}$$

and

$$\partial\pi_2/\partial x_2 \equiv -Kp_2(x_1, x_2) - 1 = 0. \tag{5}$$

The assumptions (3) also ensure that an increase in each player's effort level harms the other, and therefore makes it strategically desirable for each to precommit his effort level in such a way as to induce a lower effort from the other in response. Whether this means a commitment at a higher or a lower level of one's own effort depends on whether the other's best response function has a negative or positive slope. The general significance of the slopes of reaction functions is discussed by Drew Fudenberg and Jean Tirole, 1984, Jeremy Bulow et al., 1985, and Jonathan Eaton and Gene Grossman, 1986.

Formally, if x_1 can be precommitted,[1]

$$d\pi_1/dx_1 = \partial\pi_1/\partial x_1 + \partial\pi_1/\partial x_2 \cdot dx_2/dx_1,$$

where the first term on the right-hand side is zero in Nash equilibrium, $\partial\pi_1/\partial x_2 = Kp_2$ is negative, and dx_2/dx_1 is the slope of player 2's best response function. Differentiating along his first-order condition (5), we have

$$\frac{dx_2}{dx_1} = -\frac{\partial^2\pi_2/\partial x_2\partial x_1}{\partial^2\pi_2/\partial x_2^2} = \frac{-Kp_{12}}{-\partial^2\pi_2/\partial x_2^2}. \tag{6}$$

The denominator is positive by the second-order condition. Then

$$\frac{d\pi_1}{dx_1} = -\frac{\partial\pi_1}{\partial x_2} \cdot \frac{Kp_{12}}{-\partial^2\pi_2/\partial x_2^2}. \tag{7}$$

This is positive, giving player 1 an incentive to overcommit to effort, if and only if p_{12} is *positive*.

Likewise, if player 2 can precommit,

$$\frac{dx_1}{dx_2} = -\frac{\partial^2\pi_1/\partial x_1\partial x_2}{\partial^2\pi_1/\partial x_1^2} = \frac{Kp_{12}}{-\partial^2\pi_1/\partial x_1^2} \tag{8}$$

and

$$\frac{d\pi_2}{dx_2} = \frac{\partial\pi_2}{\partial x_1} \cdot \frac{Kp_{12}}{-\partial^2\pi_2/\partial x_2^2}. \tag{9}$$

Therefore, player 2 has an incentive to overcommit to effort if and only if p_{12} is *negative*.

We see that the two players' strategic incentives are necessarily in opposite directions, except when $p_{12} = 0$, in which case neither has a strategic incentive for precommitment to effort locally away from the Nash level. I will show that the latter must occur when there is perfect symmetry, and then explore the consequences of asymmetry.

Perfect symmetry between the players means that if their effort levels were interchanged, so would the probabilities. Therefore,

$$p(x_2, x_1) = 1 - p(x_1, x_2). \tag{10}$$

Differentiating with respect to x_1,

$$p_2(x_2, x_1) = -p_1(x_1, x_2). \tag{11}$$

[1] The same qualitative results are obtained if we introduce an indirect instrument y_1 for precommitment. Of course a separate variable is essential if one wants to study a two-stage Nash game when both players use precommitment.

In particular, for any x,

$$p_2(x,x) = -p_1(x,x).$$

Looking at the first-order conditions (4), (5), we see that under the usual mild restrictions there is a symmetric Nash equilibrium, defined by $Kp_1(x,x)=1$. Now differentiate (11) with respect to x_2:

$$p_{21}(x_2,x_1) = -p_{12}(x_1,x_2).$$

Therefore, for any x,

$$p_{21}(x,x) = -p_{12}(x,x).$$

But

$$p_{21}(x,x) = p_{12}(x,x).$$

So both must be zero in the symmetric Nash equilibrium. This proves the result.

Without perfect symmetry, little can be said about p_{12} in general. Therefore, I will consider special functional forms. There are two natural possibilities. The first is the logit form used in models of discrete choice; see Daniel McFadden, 1973. It is used for contests by Rosen, 1986, and Mortensen, 1982, and in a special case by Tullock, 1980. It also arises in a modified way (because of discounting) in models of patent races; see Loury, 1979, and Dasgupta and Stiglitz, 1980. We have

$$(12)\quad p(x_1,x_2) = \frac{f_1(x_1)}{f_1(x_1)+f_2(x_2)},$$

where f_1 and f_2 are increasing functions. It is a routine matter to verify that

$$(13)\quad p_{12} = f_1'f_2'(f_1-f_2)/(f_1+f_2)^3.$$

Therefore, $p_{12}>0$ if and only if $f_1>f_2$, that is, $p>1/2$.

The second special form, the probit, arises in some of the literature on contests as incentive schemes. Suppose efforts x_1 and x_2 yield sure components of outcomes $f_1(x_1)$ and $f_2(x_2)$,[2] but the full outcomes also contain noise components ε_1 and ε_2. Then player 1 wins if and only if

$$f_1(x_1)+\varepsilon_1 > f_2(x_2)+\varepsilon_2$$

or

$$\varepsilon_2-\varepsilon_1 < f_1(x_1)-f_2(x_2).$$

Let G be the cumulative distribution function of $(\varepsilon_2-\varepsilon_1)$. Then

$$(14)\quad p(x_1,x_2) = G(f_1(x_1)-f_2(x_2)).$$

Differentiating this twice, we have

$$(15)\quad p_{12} = -g'(f_1(x_1)-f_2(x_2)) \times f_1'(x_1)f_2'(x_2),$$

where g is the density corresponding to G. If the noise effects are symmetric between players, we have g symmetric about zero. Assume it is unimodal, as will usually be the case. Then $p_{12}>0$ if and only if $f_1(x_1)-f_2(x_2)>0$, that is, $p>1/2$.

For both specifications, we have found that the player favored to win is the one who has the strategic incentive to overexert, and the underdog, to ease up. The intuition behind this result is explained in Figure 1. This is based on the logit model; the probit case can be explained similarly. In panel (a) of Figure 1, the curve TT shows the total gross return to player 2, $[1-p(x_1,x_2)]K$, as a function of his effort x_2, for a fixed level of x_1. In panel (b), the curve MM is the corresponding marginal. A slight increase in x_1 shifts these curves to $T'T'$ and $M'M'$, respectively. The marginals cross at the point P. To its left, an increase in x_1 reduces the marginal return to x_2, which leads 2 to reduce x_2 if x_1 is precommitted. If player 1 is the favorite, $(1-p)$ is low, and x_2 lies to the left of P. Therefore, it is in 1's strategic interest

[2] Nalebuff and Stiglitz (1983) and several others take what I call $f_1(x_1)$ and $f_2(x_2)$ as the choice variables; then asymmetries show up in the costs of effort for the two players. The two methods are, of course, formally equivalent. In econometric uses of the probit, ε_1 and ε_2 are assumed to be normally distributed. Weaker assumptions stated later suffice for my purpose.

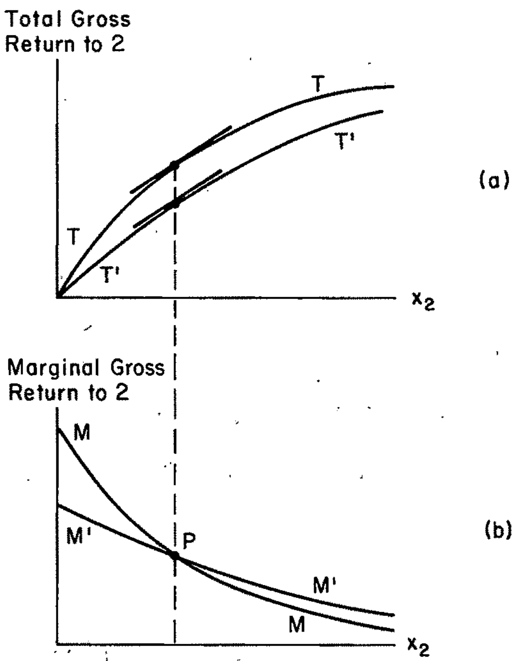


FIGURE 1

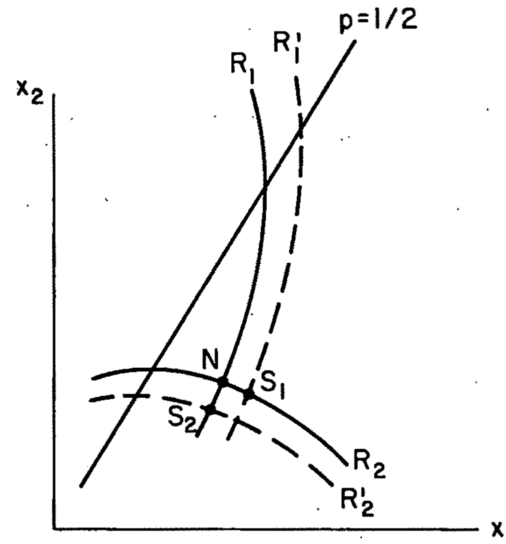


FIGURE 2

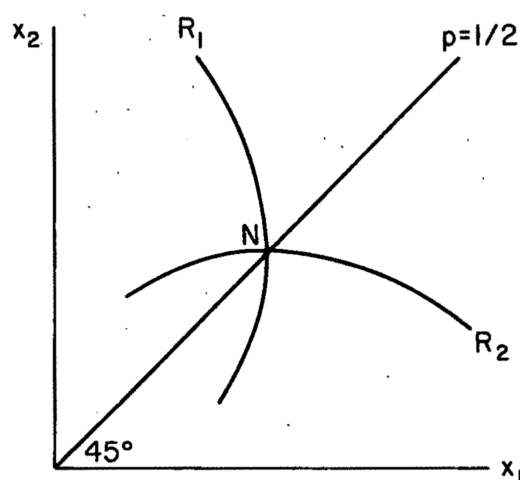


FIGURE 3

to commit to overexertion. The opposite holds if player 1 is the underdog.

Figure 2 develops the corresponding intuition using "reaction" or "best response" functions. Player 1 is the favorite. In a neighborhood of the Nash equilibrium N, the favorite's best response function (R_1) is upward sloping and that of the underdog (R_2) is downward sloping. Then player 1 has the incentive to make a strategic precommitment to a higher x_1, moving the outcome from N to S_1. Player 2 gains if he can commit to a lower x_2, and move the outcome from N to S_2. When precommitment is made by means of a separate variable y_1 (respectively, y_2), the effect is to shift the best response function R_1 to R_1' (respectively, R_2 to R_2'), thereby achieving the desired shift of the outcome. If both players can precommit to variables y_1 and y_2, in the perfect equilibrium of the two-stage game there is a tendency to shift R_1 and R_2 in the directions indicated. Depending on the relative magnitudes of the shifts, *either* x_1 will be higher and x_2 lower than at N, *or* both will be lower.

The perfectly symmetric case is shown in Figure 3. Here R_1 is vertical and R_2 horizontal at the Nash equilibrium. Therefore, there is no local incentive for strategic manipulation of effort. The possibility of gain from large shifts depends on the relative *curvatures* of the other player's best response function and one's own payoff contours. That matter is going to be too context-specific to be settled theoretically.

II. Many Players

Index the players $i=1,2,\ldots,n$, and suppose player 1 can precommit his effort level. I choose the logit specification for the probability of a win for player i:

$$p_i = f_i(x_i)/S, \quad i=1,2,\ldots,n, \tag{16}$$

where

$$S = \sum_{j=1}^{n} f_j(x_j). \tag{17}$$

The functions f_i are assumed to be increasing, and concave. The latter is sufficient for second-order conditions.

The payoffs are given by

$$\pi_i = Kp_i - x_i, \qquad i=1,2,\ldots,n. \tag{18}$$

Given x_1, the remaining players achieve a Nash equilibrium in their effort levels. The first-order conditions are

$$\partial\pi_i/\partial x_i = K\,\partial p_i/\partial x_i - 1 = 0, \qquad i=2,3,\ldots,n. \tag{19}$$

These define $(x_2, x_3,\ldots,x_n)$ as functions of x_1. Then player 1's calculation of the payoff to precommitment is

$$\frac{d\pi_1}{dx_1} = \frac{\partial\pi_1}{\partial x_1} + \sum_{j=2}^{n} \frac{\partial\pi_1}{\partial x_j}\frac{dx_j}{dx_1}. \tag{20}$$

The first term is zero in the full Nash equilibrium of all players. The other terms give the strategic effect of manipulating the other players' effort levels.

Introduce the notation, for $i=1,2,\ldots,n$,

$$q_i = f_i'(x_i)/S, \qquad r_i = f_i''(x_i)/S \tag{21}$$

Then it is routine to verify that

$$\partial p_i/\partial x_i = q_i(1-p_i), \qquad \partial^2 p_i/\partial x_i^2 = (1-p_i)(r_i - 2q_i^2) \tag{22}$$

and for $j \neq i$

$$\partial p_i/\partial x_j = -p_i q_j, \qquad \partial^2 p_i/\partial x_i \partial x_j = -q_i q_j (1-2p_i). \tag{23}$$

Using (23), we can write (20) as

$$\frac{d\pi_1}{dx_1} = -Kp_i \sum_{j=2}^{n} q_j \frac{dx_j}{dx_1}. \tag{24}$$

To find the right-hand side, differentiate the first-order conditions (19) totally, using (22) and (23). This gives, for $i=2,3,\ldots,n$,

$$(1-p_i)(r_i - 2q_i^2)dx_i - q_i(1-2p_i)\sum_{j\neq i} q_j dx_j = 0$$

or

$$\left[(1-p_i)(r_i-2q_i^2) + q_i^2(1-2p_i)\right]dx_i - q_i(1-2p_i)\sum_{j=1}^{n} q_j dx_j = 0$$

or

$$q_i dx_i = -\theta_i \sum_{j=1}^{n} q_j dx_j, \tag{25}$$

where

$$\theta_i = -\frac{q_i^2(1-2p_i)}{(1-p_i)r_i - q_i^2}. \tag{26}$$

Then

$$\sum_{i=2}^{n} q_i dx_i = -\left\{\sum_{i=2}^{n}\theta_i\right\}\left\{\sum_{j=1}^{n} q_j dx_j\right\}$$

or

$$\left\{\sum_{i=2}^{n} q_i dx_i\right\}\left\{1+\sum_{i=2}^{n}\theta_i\right\} = -\left\{\sum_{i=2}^{n}\theta_i\right\} q_1 dx_1$$

or

$$(27)\quad \sum_{i=2}^{n} q_i dx_i = -\left\{\sum_{i=2}^{n}\theta_i\right\} q_1 dx_1 \Big/ \left\{1+\sum_{i=2}^{n}\theta_i\right\}.$$

Substituting in (24), we finally have

$$(28)\quad d\pi_1/dx_1 = Kp_1q_1\left\{\sum_{i=2}^{n}\theta_i\right\}\Big/\left\{1+\sum_{i=2}^{n}\theta_i\right\}.$$

To determine the sign of this, examine the definition (26) of θ_i. Each denominator is negative since each r_i is nonpositive. If each $p_i < 1/2$, then each numerator is positive. Then each θ_i is positive, and so is $d\pi_1/dx_1$. Thus the absence of an odds-on favorite among the remaining players is a sufficient condition for commitment to overexertion by player 1. In particular, the favorite is sure to overexert given an opportunity to precommit.

If $n = 2$, the right-hand side becomes

$$Kp_1q_1\theta_2/(1+\theta_2) = -Kp_1q_1q_2^2(1-2p_2)/\{(1-p_2)(r_2-2q_2^2)\},$$

and $p_2 < 1/2$ becomes necessary as well as sufficient for $d\pi_1/dx_1 > 0$. Thus the model contains the analysis of Section II as a special case, as of course it should.

III. Applications and Extensions

The two-person result of Section I finds support in many sports events where one player or team is clearly favored to win. Such a player, or the manager of such a team, often declares that "since he is expected to win, he must try all the harder." The underdog, on the other hand, is "under no pressure, and is just going to enjoy the occasion." There is some objective truth to these assertions, but to a considerable extent the pressure on a favorite is of his own making, as is the relaxed mood of an underdog. Therefore, these can be seen as ways of altering one's own preferences in advance, that is, as instruments of commitment that will credibly alter future behavior.

The result runs counter to the popular belief that No. 2's "try harder." But that statement never clarifies "than what or whom." My comparison is between the same player's uncommitted and precommitted effort levels. An underdog may be making more effort than the favorite; $p > 1/2$ is compatible with $x_1 < x_2$ if the two players' functions f_1 and f_2 are suitably different. It is also believed that underdogs use riskier strategies, or that favorites play too safely, but such questions need a different model.

Next consider rent seeking. Suppose a contract or an import license is being awarded. The two contenders are not equally efficient, and the bureaucrats or the politicians pay *some* attention to this aspect. Let x_1, x_2 be the bribes offered by the two, and let $p(x_1, x_2)$ be *either* the probability that player 1 gets the whole contract (license) *or* his share. Then the result says that the more efficient user, if given first access to the bureaucrat or the politician, will bribe more than he would in the case of simultaneous access.

One case of oligopoly fits the model directly, namely a homogeneous product with unit-elastic demand. Then we can interpret K as the total market revenue and $p(x_1, x_2)$ as firm 1's market share.

Patent races are more natural settings, especially in the international arena where countries make commitments to *R&D* programs. If one thinks of the U.S.-Japan rivalry in this way, one would use the above result to conclude that since the United States was favored over Japan in the late 1970's, therefore Japan should have committed to underexertion. It clearly did not. However, an explanation may be found in the model of Section II. The United States was not a single player, but a group of separate firms in which no one was an odds-on favorite.

Discounting influences the result in the same direction. In the model of Loury, 1979, and Dasgupta and Stiglitz, 1980, the research effort x_i of firm i is expected at time 0, and buys a conditional probability of success (hazard rate) $f_i(x_i)$ in a Poisson process. Let

K be the capitalized royalties from the innovation. Then we find

$$\pi_i = Kf_i(x_i)/S - x_i \tag{29}$$

where

$$S = f_1(x_1) + f_2(x_2) + r, \tag{30}$$

r being the discount rate. Now

$$\frac{\partial^2 \pi_1}{\partial x_1 \partial x_2} = \frac{Kf_1' f_2(f_1 - f_2 - r)}{(f_1 + f_2 + r)^3} \tag{31i}$$

and

$$\frac{\partial^2 \pi_2}{\partial x_1 \partial x_2} = \frac{Kf_1' f_2'(f_2 - f_1 - r)}{(f_1 + f_2 + r)^3}. \tag{31ii}$$

Compare these with (13). The introduction of discounting makes both of these negative in more cases, and therefore favors a commitment to overexertion by both players. Mathematically, the discount rate is just like having a third player whose response function happens to be totally inelastic.

In conclusion, I should point out some issues that were not addressed here. The commitments considered were unconditional, that is, they amounted to a first-mover advantage. Conditional commitments can achieve more. In a sense they can achieve too much, for player 1 can drive player 2 out of the game by committing himself to choosing a destructively large x_1 if x_2 is positive. As usual, the crucial question is the credibility of such conditional commitments, and it is better studied in a context-specific way.

Another question is that of efficiency of the outcome. In rent seeking, it may be socially desirable to have the minimum total investment. Then the underdog briber should go first. In *R&D* races, we may be concerned with the industry's total expected profit, or may want to add on some spillover benefit to society. In both cases, it is of interest to compare the Nash equilibrium with a leader-follower outcome. But once again, the general answers are so vague that the question is best addressed specifically in each application.

REFERENCES

Bulow, Jeremy, Geanakoplos, John and Klemperer, Paul, "Multimarket Oligopoly: Strategic Substitutes and Complements," *Journal of Political Economy*, June 1985, *93*, 488–511.

Dasgupta, Partha and Stiglitz, Joseph, "Uncertainty, Industrial Structure, and the Speed of *R&D*," *Bell Journal of Economics*, Spring 1980, *11*, 1–28.

Eaton, Jonathan and Grossman, Gene, "Optimal Trade and Industrial Policy Under Oligopoly," *Quarterly Journal of Economics*, May 1986, *101*, 383–406.

Fudenberg, Drew and Tirole, Jean, "The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look," *American Economic Review Proceedings*, May 1984, *74*, 361–66.

Green, Jerry and Stokey, Nancy, "A Comparison of Tournaments and Contracts," *Journal of Political Economy*, June 1983, *91*, 349–64.

Holmstrom, Bengt, "Moral Hazard in Teams," *Bell Journal of Economics*, Autumn 1982, *13*, 324–40.

Lazear, Edward and Rosen, Sherwin, "Rank Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, October 1981, *89*, 841–64.

Lee, Tom and Wilde, Louis, "Market Structure and Innovation: A Reformulation," *Quarterly Journal of Economics*, March 1980, *94*, 429–36.

Loury, Glenn, "Market Structure and Innovation," *Quarterly Journal of Economics*, August 1979, *93*, 395–410.

McFadden, Daniel, "Conditional Logit Analysis of Qualitative Choice Behavior," in Paul Zarembka, ed., *Frontiers in Econometrics*, New York: Academic Press, 1973, 105–42.

Mortensen, Dale T., "Property Rights and Efficiency in Mating, Racing, and Related Games," *American Economic Review*, December 1982, *72*, 968–79.

Nalebuff, Barry and Riley, John, "Asymmetric Equilibria in the War of Attrition," *Journal of Theoretical Biology*, July 1985, *113*, 517–27.

______, and Stiglitz, Joseph, "Prizes and

Incentives: Toward a General Theory of Compensation and Competition," *Bell Journal of Economics*, Spring 1983, *14*, 21–43.

Riley, John, "Evolutionary Equilibrium Strategies," *Journal of Theoretical Biology*, January 1979, *76*, 109–23.

Rosen, Sherwin, "Prizes and Incentives in Elimination Tournaments," *American Economic Review*, September 1986, *76*, 701–15.

Tullock, Gordon, "Efficient Rent Seeking," in J. Buchanan, R. Tollison, and G. Tullock, eds., *Toward a Theory of the Rent-Seeking Society*, College Station: Texas A&M University Press, 1980.

Priority Service: Pricing, Investment, and Market Organization

By HUNG-PO CHAO AND ROBERT WILSON*

Priority service offers a menu of contingent contracts for distribution of scarce supplies. Prices inducing customers' efficient self-selection are expectations of spot prices for comparable service. Customers' selections reveal the benefit of capacity expansion. Priority service can be implemented via sale of "priority points" or via provision of compensatory insurance. Several priority classes suffice to obtain most efficiency gains. Priority service Pareto dominates random rationing if excess revenue is refunded equally to customers.

Priority service refers to an array of contingent forward delivery contracts offered by a seller. Each customer's selection of one contract from the menu determines the customer's service order or priority. In each contingency, the seller rations supplies by serving customers in order of their selected priorities until the supply is exhausted or all customers are served.

Various forms of priority service are widely used, but it has not received much attention in the economic literature until recently. Priority service is closely related to three distinct types of economic activity: 1) product differentiation, 2) rationing, and 3) spot and future markets.

First, priority service can be viewed as a special form of product differentiation in which the market is segmented into a spectrum of priority classes. Those customers willing to pay higher prices are assigned higher priority in receiving the product or service. The importance of this scheme is underscored by Milton Harris and Arthur Raviv (1981), who find that among all incentive mechanisms, priority service allows a monopoly to extract the highest profits from selling a scarce supply. In the more specific context of electric power, Maurice Marchand (1974) and John Tschirhart and Frank Jen (1979) consider a similar pricing scheme in which interruptible service is priced according to service reliability. Interruptible service for power, implemented on a large scale, offers substantial improvements in economic efficiency compared to the indiscriminate rationing methods usually employed (Shmuel Oren et al., 1986).

Second, priority service can be seen as a rationing scheme for curtailing excess demand in the event of deficient supply, where both efficiency and equity are relevant considerations. The theory of efficient rationing suggests that allocation should be according to customers' valuations of service. Because customers' selections reflect their valuations, priority service implements precisely this method of allocation. Moreover, we show in Section IV that compared to a random rationing scheme with a fixed price, one implementation for priority service is Pareto superior. This is because the efficiency gains that are realized can be distributed to increase every customer's expected net benefit, without affecting customers' incentives to self-select efficiently.

Third, priority service can be interpreted as a form of market organization that supplements, and in some cases supplants, spot markets. Spot prices are revised continually, whereas priority service contracts cover a period of extended duration. In principle, the price charged for each priority class is the expectation of what the spot prices would be for the same quality of service purchased in the spot markets. Corresponding to each priority class is an imputed reservation price

*Electric Power Research Institute, Palo Alto, CA 94303, and Stanford University, Stanford, CA 94305, respectively. We are grateful to Shmuel Oren and Stephen Smith, with whom we have worked jointly on this topic for several years. The National Science Foundation supported part of this work under grant no. SES-8605666. The views expressed here are our own.

that is the maximum spot price at which the customer would make spot purchases. Priority service can be a less costly form of market organization if supplies are nonstorable, customers' valuations are stable over time, and transaction costs are significant.

Besides its theoretical justification as an efficient form of rationing, the construction of priority service contracts in terms of service orders has practical aspects. One can argue that delivery under forward contracts is inherently contingent on sufficiently low spot prices. Such contracts typically promise delivery at a future time in specified contingencies (or uncontingently). Suppose that spot markets operate continually for a nonstorable product or service. In this case, because spot prices have no natural upper bounds, one anticipates that the seller who draws supplies from the spot market will fulfill a forward contract only if the spot price is below some critical level determined by the cost of breaching the contract. Similarly, even if delivery is contingent on the spot price, the buyer will stipulate delivery only if the spot price is low enough. Therefore, aspects of priority service typically are embedded within any auxiliary futures contracts and other contingent forward contracts. More complex contracts can be written most simply and efficiently by relying on priority assignments or equivalent clauses specifying reservation prices.

Compared with spot pricing, priority service offers two major advantages. One is that it yields important information about the distribution of customers' valuations that can be used to guide capacity planning. This information is unavailable from the observed choice behavior of customers in a spot market. That is, a spot market is essentially an algorithm to determine an efficient allocation in a particular contingency, namely, the particular maximum reservation price to be served in that contingency. In contrast, the process of self-selection among priority service contracts enables the seller to infer the allocation rule for every contingency. As we show in Section VIII, this additional information enables a calculation of customers' willingness-to-pay for the higher service reliability provided by additional capacity.

A second major advantage of priority service is that it enables supplemental insurance provisions to be incorporated into the contracts. When the role of customers' risk aversion is recognized, efficient risk sharing requires that any form of market organization be accompanied by insurance provisions. If, as seems realistic, the producer or a third-party underwriter is the most efficient bearer of risk, then the efficient insurance contracts cover all or most of the customers' risk. It is therefore in the underwriter's interest to allocate supplies so as to minimize claims. In Section V, we formalize this argument and show that with perfect insurance (i.e., offered at actuarially fair premia) the efficient incentive scheme entails allocation of supplies according to customers' valuations of service. Thus, inclusion of insurance provisions necessitates a version of priority service embedded within the insurance contracts.

The electric power industry represents an ideal candidate for implementing priority service. In part, this is because recent advances in the microelectronic technologies of metering, control, and communication have made it feasible. The principal alternative is spot pricing. Although spot pricing is used in wholesale markets for bulk trades among power producers, proposals to use spot pricing in retail markets have not been successful. A common explanation for the rejection of spot pricing in retail markets is that customers want prior assurance about what their monthly bills will be. But there are several more fundamental reasons.

First, there is the ultimate argument of infeasibility. Electricity is essentially nonstorable, whereas failures of generating equipment can occur within time frames of milliseconds. (Spot prices can therefore fluctuate quickly and greatly.) This requires allocation rules that can be implemented comparably quickly. Technologies that enable such quick responses by customers are not feasible presently. Moreover, there are no known methods for establishing new equilibrium prices without time-consuming iterations or collection of bids and offers.

A second argument, appropriate to intermediate time frames, introduces transaction costs. Continually monitoring spot prices,

and adjusting demands responsively, imposes appreciable costs on customers. Even if predetermined response rules are implemented automatically, a considerable investment in equipment is required. But this is essentially what priority service does: it takes advantage of the fact that the optimal rationing rule is based on priority assignments determined by customers' relative valuations of service. And, it exploits the empirical evidence that customers' valuations of service are persistent, or serially correlated to a high degree. This makes automation of rationing rules feasible, and centralization of the rationing rules economizes on the costs of implementation.

In Section VII we strengthen this argument. We show that priority service with only a few priority classes (and random rationing within each priority class) realizes most of the potential gains from efficient rationing. Thus, the fine differentiation of spot prices that is necessary to balance demand and supply continually is not essential to attainment of efficiency, given that some transaction costs are associated with either market organization. In contrast to spot pricing, priority service leads to a market organization in which only a relatively few standardized contracts are traded. These contracts supplant the implicitly infinite variety of spot prices, and the continual intertemporal variation, with only slight efficiency losses and appreciable savings on the costs of implementation.

The objective of this paper is to investigate the optimal pricing and investment rules for priority service, as well as alternative market organizations required for its implementation. For concreteness, we draw on the case of the electric power industry in the characterization of the production technology. The formulation is otherwise rather general.

Our analysis of priority service incorporates three important considerations. First, in the short term with a fixed capacity configuration and demand characteristics, the relative prices for the various service priorities are adjusted to provide incentives to consumers to self-select according to their willingness to pay. This results in a balanced distribution of selections with sufficient low-priority demand to provide more reliable service to high-priority demand. Second, in the long term, integration with capacity planning allows the priority charges to reflect the associated capacity and operating costs. Third, implementations of priority service can employ several alternative forms of market organization. The importance of market organization for priority service is emphasized by William Vickrey (1971), who enumerates a variety of possibilities in several industries. Our aim here is to establish the theoretical properties of several candidate forms of market organization.

The sections of the paper are organized as follows. Section I develops the basic model that characterizes the consumers' choice behavior and the cost structure of the underlying production technology. Section II addresses the design of an optimal menu of priority service options. Section III compares the welfare implications of priority service and spot pricing. Section IV compares priority service with the random rationing scheme. Section V develops an extension of the basic model to include the case that consumers may be risk averse to the financial consequences of interruption. Section VI provides an example that illustrates several of the results. Section VII shows the asymptotic performance of priority service when there will be a finite number of priority classes. Section VIII addresses the issue of optimal capacity expansion under priority service. Section IX describes three implementations of priority service that differ in terms of the contract forms and market organization. In Section X, we conclude with a brief summary.

I. The Basic Model

In this section we describe the two main components of our formulation: a consumer choice model and a cost model. Our approach is one of static partial equilibrium. Income effects are omitted, and the consideration of risk aversion is deferred until Section V. The representation of uncertainty is quite general. Throughout this paper, all random variables are represented as a function of ω on Ω, an abstract sample space associated with a σ field, and a probability mea-

sure. The random outcome ω is not fully observable in general.

We consider a pricing scheme characterized by a menu $M = \{(p,s,r)\}$ of options. For each option (p,s,r), p is the priority charge (payable in advance), s is the service charge (payable as service is provided), and r is the service reliability, which is the probability of receiving the product or service. Selection of an option by a consumer is equivalent to acceptance of a contingent forward contract for delivery in an event having the specified probability.

Our pricing scheme can be viewed as a generalized version of the priority pricing scheme studied by Harris-Raviv (1981), for it includes as a special case the single-price scheme when the price menu contains only two options, one of which is the default option $(0,0,0)$ of not buying. For convenience hereafter, we refer to such a menu as a priority service scheme.

A. *The Consumer Choice Model*

The model has an asymmetric information structure. We assume that each consumer's willingness-to-pay for service is privately known. Each consumer can freely choose from the menu any priority option and assign it to any increment of his consumption. Therefore, without loss of generality, each consumer can be simply characterized by a single unit of demand and the associated marginal willingness-to-pay v, which takes a value in the interval from 0 to V. The aggregate demand, or the total number of demand units for each type, is uncertain. The aggregate demand function is represented by $D(\cdot,\omega)$, and the inverse demand function, or the willingness-to-pay function, is represented by $P(\cdot,\omega)$, both contingent on the state ω.

The objective of each consumer is to choose from the menu M a priority option that maximizes his expected surplus. Therefore, for each v, the consumer's problem is to solve

$$(1)\quad S(v) = \max\{r(v-s)-p \mid (p,s,r)\in M\}.$$

We assume that the expected total charge is always nonnegative, that is, $p+rs\geq 0$ for every $(p,s,r)\in M$. This implies that $S(0)=0$. Let $(p(v),\ s(v),\ r(v))$ denote an optimal solution to (1). The optimal consumer behavior can be characterized as follows.

THEOREM 1: *The optimal consumer choices satisfy the following conditions:* (A) *$r(v)$ is nondecreasing in v;* (B) *$p(v)+r(v)s(v)$ is nondecreasing in v; and* (C) *$p(v)+r(v)s(v) = \int_0^v [r(v)-r(u)]\,du$, for every v.*

PROOF:

(A) By definition $(p(v),\ s(v),\ r(v))$ maximizes the expected surplus of a consumer with a willingness-to-pay v; hence for any u,

$$(2)\quad \begin{aligned} S(v) &= r(v)(v-s(v))-p(v) \\ &\geq r(u)(v-s(u))-p(u) \\ &= S(u)+r(u)(v-u) \end{aligned}$$

or

$$(3)\quad S(v)-S(u)\geq r(u)(v-u).$$

Switching the role of u and v in (3) yields

$$(4)\quad S(u)-S(v)\geq r(v)(u-v).$$

Combining (3) and (4) we obtain

$$(5)\quad \begin{aligned} r(v)(v-u) &\geq S(v)-S(u) \\ &\geq r(u)(v-u). \end{aligned}$$

It follows from (5) that if $v>u$ then $r(v)\geq r(u)$, that is, $r(v)$ is nondecreasing in v.

(B) Suppose $v>u$ and thus $r(v)\geq r(u)$. Then

$$(6)\quad \begin{aligned} & r(v)(u-s(v))-p(v) \\ &\leq S(u) \\ &= r(u)(u-s(u))-p(u) \\ &\leq r(v)u-r(u)s(u)-p(u). \end{aligned}$$

It follows that $r(v)s(v)+p(v)\geq r(u)s(u)+p(u)$.

(C) Since $r(\cdot)$ is monotone, it is continuous almost everywhere. Dividing (5) by $(v-u)$ and then passing to the limit as $u\rightarrow v$

yields

$$S'(v) = r(v). \tag{7}$$

Integrating (7) with the initial condition that $S(0) = 0$ yields the desired result.

Theorem 1 can be interpreted in more intuitive terms as follows. Conditions (A) and (B) require that consumers with higher willingness-to-pay select more reliable plans and pay more. In other words, consumers reveal their preferences through their choices, and therefore, with this information, the output can be rationed more efficiently. Note that the left side of condition (C) represents the expected total charge associated with the priority chosen by a type v consumer. Condition (C) indicates that the expected total charge should be sufficient to compensate the losses incurred by those who choose lower priorities in order to provide the service reliability required for the higher priority. To see this, note that the loss incurred by a consumer with willingness-to-pay u due to a change of reliability $dr(u)$ is given by $u\ dr(u)$. Integrating this term for all consumers who are willing to pay less than v for the service yields

$$\int_0^v u\, dr(u) = vr(v) - \int_0^v r(u)\, du,$$

which is precisely the right side of condition (C).

The converse of Theorem 1 can be stated as follows.

THEOREM 2: *If* $p(\cdot)$, $s(\cdot)$, *and* $r(\cdot)$ *satisfy conditions* (A) *and* (C) *stated in Theorem* 1, *and* $M = \{(p(v), s(v), r(v)) | 0 \le v \le V\}$, *then for each* v, $(p(v), s(v), r(v))$ *is an optimal solution to problem* (1).

PROOF:

For any v and v' in $[0, V]$, we have

$$\begin{aligned} r(v')(v - s(v')) - p(v') &= r(v')v - \int_0^{v'} [r(v') - r(u)]\, du \\ &= \int_0^v r(u)\, du - \int_v^{v'} [r(v') - r(u)]\, du. \end{aligned} \tag{8}$$

Since $r(\cdot)$ is nondecreasing, the last term in (8) is always nonnegative: if $v < v'$, then the integrand is nonnegative; or if $v > v'$, then the integrand is nonpositive but the direction of integration is backward. Therefore, we obtain

$$\begin{aligned} r(v)(v - s(v)) - p(v) &= \int_0^v r(u)\, du \\ &\ge r(v')(v - s(v')) - p(v'). \end{aligned} \tag{9}$$

Here, the equality holds if and only if $r(v) = r(v')$.

Theorem 2 implies that we can construct a price menu that will induce optimal consumer choices consistent with any prespecified allocation of service reliability that satisfies condition (A). That is, given any nondecreasing mapping $r(\cdot)$ from $[0, V]$ to $[0, 1]$, we find $p(\cdot)$ and $s(\cdot)$ so that condition (C) holds. Theorem 2 indicates that a consumer with willingness-to-pay v will choose precisely the option $(p(v), s(v), r(v))$ from the price menu M. Therefore, through consumers' self-selection, priority service provides important information about the distribution of consumers' willingness to pay.

B. *The Cost Model*

For concreteness, we describe the production technology in the context of electric power generation. There are two notable features: the supply uncertainty due to random outages of power plants, and a nonlinear cost structure resulting from multiple technologies. Specifically, we assume that there are n technologies with marginal capacity costs k_i, and marginal operating costs c_i for $i = 1, \ldots, n$, respectively. These technologies are ranked in ascending order by operating cost. For each technology, the total capacity consists of a continuum of homogeneous generation units, each of which is subject to random failures. Let X_i denote the installed capacity of technology i, and the random function $Y_i(X_i, \omega)$ denotes the available capacity, whose realization is known to the

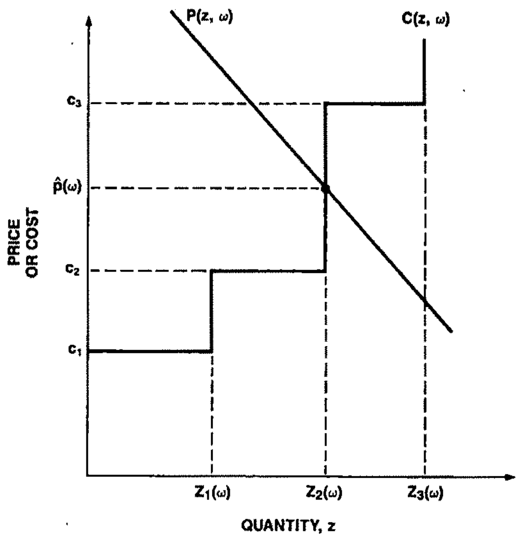


FIGURE 1. DETERMINATION OF SPOT PRICE

producer. The unit availability factor is denoted by a constant $a_i = E\{\partial Y_i(X,\omega)/\partial X\}$. We assume here that the capacity increments, that is, $Y_i(dX,\omega)$, are independent of each other and of all other random variables.

The system operation is based on the prespecified loading order of technologies from 1 to n. For a given capacity configuration $(X_1, X_2, \dots, X_n)$, the total available capacity of technologies $1,\dots,i$ is denoted by

$$(10) \qquad Z_i(\omega) = \sum_{j=1}^{i} Y_j(X_j,\omega).$$

As illustrated in Figure 1, the short-run marginal cost function is given by

$$(11) \quad C(z,\omega) = c_i, \text{ if } Z_{i-1}(\omega) < z \le Z_i(\omega).$$

II. Optimal Price Menu

This section is concerned with the design of a price menu that maximizes the social welfare. We proceed in two steps. First, we present the conditions for optimal allocation, assuming that ω is fully observable. Then assuming only that the utility knows the *distribution* of ω, we construct a price menu and demonstrate that it will induce consumer choices consistent with the optimal allocation obtained with perfect information.

Generation dispatch and load control are two principal operations by which an electric utility balances the demand and supply of power in real time. Suppose that the utility has perfect knowledge about consumers' preferences. Then the conditions for optimal generation dispatch and load control can be summarized as: 1) generation units are dispatched in ascending order of the unit operating cost; 2) consumers are served in descending order of their willingness-to-pay; 3) these operations are continued until the marginal operating cost exceeds the marginal willingness-to-pay. In other words, the problem is to find the intersection of the marginal willingness-to-pay function and the marginal cost function, as illustrated in Figure 1. Denote by $\hat{p}(\omega)$ the instantaneous equilibrium price, or the spot price, associated with a given random outcome ω. Then the service reliability of a type v consumer can be expressed as

$$(12) \qquad R(v) = \Pr\{\hat{p}(\omega) \le v\}.$$

This indicates that the consumer is served in the events for which the spot price is less than his willingness to pay. Formally, the spot price can be characterized as follows:

$$(13) \quad \hat{p}(\omega)$$

$$= \min\{\max[P(z,\omega), C(z,\omega)] \,|\, z \ge 0\}$$

$$= \min\{\max[P(Z_i(\omega),\omega), c_i] \,|\, i+1,\dots,n\}.$$

By inspecting Figure 1, we obtain the following alternative characterization of spot prices, which can be derived from (13) in a straightforward manner.

LEMMA 1: *The following relations hold*: (i) $\hat{p}(\omega) = P(Z_i(\omega),\omega)$, iff $c_i \le P(Z_i(\omega),\omega) \le c_{i+1}$; (ii) $\hat{p}(\omega) = c_i$, iff $P(Z_i(\omega),\omega) \le c_i \le P(Z_{i-1}(\omega),\omega)$, *where* $i = 1,\dots,n$ *and* $c_{n+1} \underline{\Delta} V$.

Given (12), we can construct a price menu M^* as follows:

$$(14)\qquad r^*(v)=\begin{cases} R(v), & \text{if } v\geq c_i,\\ 0, & \text{if } v<c_i.\end{cases}$$

$$(15)\qquad s^*(v)=\begin{cases} c_i, & \text{if } c_i\leq v<c_{i+1},\\ 0, & \text{if } v<c_i.\end{cases}$$

$$(16)\qquad p^*(v)=\int_0^v[r^*(v)-r^*(u)]\,du - s^*(v)r^*(v).$$

$$(17)\qquad M^*=\{(p^*(v),s^*(v), r^*(v))|0\leq v\leq V\}.$$

Since $R(v)$ is nondecreasing in v, it follows from Theorem 2 that M^* will induce optimal consumer choices consistent with the optimal allocation described above.

Now, let us consider a simple operating rule as follows: Dispatch generation units in ascending order of the unit operating cost and serve consumers in descending order of the service reliability that they choose, until the marginal operating cost first exceeds the service charge. Note that this operating rule depends only on the observable information, and that it leads to maximum profits in the short run. Furthermore, the three optimality conditions stated above are satisfied for the following reasons: Condition (1) obviously holds. Condition (2) is satisfied because $r^*(v)$ is nondecreasing in v. Condition (3) is satisfied, because in view of (15),

$$(18)\qquad c_i>v \text{ iff } c_i>s^*(v).$$

We have thus established the optimality of the menu M^*, which can be implemented in a variety of forms to be discussed in later sections.

III. A Comparison with the Spot Pricing Scheme

In this section we compare priority service pricing with spot pricing, assuming that $D(\cdot,\omega)$ and $X(\omega)$ are observable. Both schemes are motivated by the same efficiency considerations. The key difference is in the time frame. With the spot pricing scheme, the price is revised instantaneously as supply and demand conditions change. Priority service is generally offered as a forward contract over a longer period. These two pricing schemes are closely related, and indeed spot pricing can be viewed as a limiting case of priority service as the pricing period is reduced to zero. A practical difficulty with spot pricing is that the sample space may be so complex that it would be impossible to implement the spot price in every contingency. However, based on the arguments in Section II, we can obtain the following result.

PROPOSITION 1: *In the absence of transaction costs, the priority service scheme and the spot pricing scheme are equally efficient, and both can achieve the full-information socially optimal allocation.*

It warrants attention that the two pricing schemes will converge to the single-price scheme as the uncertainty diminishes. This is because, in a deterministic world, the sample space Ω contains only a single element, and thus the spot prices are reduced to a single price $\hat{p}$. Further, it follows from (12) that $R(v)$ equals 1, if $v\geq\hat{p}$, and equals 0, if $v<\hat{p}$. Then from (16), we obtain

$$(19)\qquad p^*(v)+s^*(v)r^*(v) =\begin{cases}\hat{p}, & \text{if } v\geq\hat{p},\\ 0, & \text{if } v<\hat{p}.\end{cases}$$

This is precisely what would be expected with a single price $\hat{p}$. Proposition 2 shows a more general and closer relation between the two pricing schemes.

PROPOSITION 2: *Under the assumption of risk neutrality, priority service pricing and spot pricing are equivalent from the perspective of individual consumers. The expected expenditures and the expected surplus of each consumer under the two pricing schemes are*

identical. That is,

$$(20)\quad p^*(v)+s^*(v)r^*(v) = E\{\hat{p}(\omega)I_{\{\hat{p}(\omega)\le v\}}(\omega)\},$$

and

$$(21)\quad [v-s^*(v)]r^*(v)-p^*(v) = E\{[v-\hat{p}(\omega)]I_{\{\hat{p}(\omega)\le v\}}(\omega)\},$$

where $I_{\{\hat{p}(\omega)\le v\}}(\omega)$ is an indicator function, which takes on value 1 or 0 depending on whether ω belongs to the set $\{\omega\colon \hat{p}(\omega)\le v\}$.

PROOF:

The expected expenditure of a type v consumer under spot pricing can be written

$$(22)\quad E\{\hat{p}(\omega)I_{\{\hat{p}(\omega)<v\}}(\omega)\}$$
$$=\int_0^V \Pr\{\hat{p}(\omega)I_{\{\hat{p}(\omega)\le v\}}(\omega)>u\}\,du$$
$$=\int_0^v \Pr\{v\ge\hat{p}(\omega)>u\}\,du$$
$$=\int_0^v [\Pr\{\hat{p}(\omega)\le v\} - \Pr\{\hat{p}(\omega)\le u\}]\,du$$
$$=\int_0^v [r^*(v)-r^*(u)]\,du.$$
$$=p^*(v)+s^*(v)r^*(v) \quad \text{(using equation (16))}.$$

Equation (21) follows directly from (20) and the relation that $r^*(v)=\Pr\{\hat{p}(\omega)\le v\}= E\{I_{\{\hat{p}(\omega)\le v\}}(\omega)\}$.

Equation (20) suggests that the expected total charge for priority service is equal to the expectation of what the spot prices would be in the events for which service is delivered under the selected priority. This result establishes the marginal cost basis for the priority service scheme.

From the utility's perspective, however, these two schemes are not necessarily equivalent. Loosely speaking, with spot pricing, the expected revenue equals the mathematical expectation of the product of the spot price and the spot demand. In contrast, with priority service, the expected revenue equals the product of the expected spot price and the expected spot demand. Since the spot price and the spot demand are generally correlated, the expected revenues and profits obtained with the two pricing schemes may be different. We demonstrate in the following a sufficient condition for the two pricing schemes to yield the same expected profits.

PROPOSITION 3: *If the marginal demand function, $\delta(v,\omega)$ ($\underline{\Delta}-\partial D(v,\omega)/\partial v$), is stochastically independent of aggregate demand and supply, then the expected profits under the two pricing schemes are identical.*

PROOF:

It follows from the analysis in Section II that the two pricing schemes will lead to the same level of output for every ω and to the same total operating costs. Therefore, what needs to be shown is that the expected revenues are the same under the two pricing schemes. Now, the expected revenue under the spot pricing scheme can be written

$$(23)\quad E\{\hat{p}(\omega)D(\hat{p}(\omega),\omega)\}$$
$$=E\left\{\hat{p}(\omega)\int_{\hat{p}(\omega)}^V \delta(v,\omega)\,dv\right\}$$
$$=\int_0^V E\{\hat{p}(\omega)I_{\{\hat{p}(\omega)\le v\}}(\omega)\delta(v,\omega)\}\,dv,$$
$$=\int_0^V E\{\hat{p}(\omega)I_{\{\hat{p}(\omega)\le v\}}(\omega)\}\times E\{\delta(v,\omega)\}\,dv \quad \text{(due to the independence assumption)}$$
$$=\int_0^V [p^*(v)+s^*(v)r^*(v)]\times E\{\delta(v,\omega)\}\,dv \quad \text{(using equation (20))}.$$

A special case of Proposition 3 occurs when demand is not stochastic. The basic model allows that demand can be stochastic if customers and the utility are imperfectly informed about the numbers of customers having the various possible valuations of units of service.

IV. A Comparison with the Random Rationing Scheme

On the grounds of efficiency, priority service is clearly superior to the random rationing scheme with a fixed price independent of ω. However, an important consideration in practice is concerned with distributional effects. The question is whether priority service will make some consumers worse off, although it benefits the society at large. In this section we demonstrate that under rather weak conditions, priority service is Pareto superior to the single-price service with random rationing. In other words, priority service can be implemented in such a way that it will disadvantage no consumers, while yielding the same expected profit for the utility.

We consider only the case that there is a single-generation technology with the constant marginal operating cost c. Assume that the demand is deterministic with $D(v)$ units having valuations greater than v, and that with perfect rationing, the probability that a consumer with willingness-to-pay v is served is $R(v)=\Pr\{D(v)\leq Z(\omega)\}$. In this section we add the technical assumption that $D(\cdot)$ and $R(\cdot)$ are differentiable.

With the single price s and random rationing, each consumer is served with the probability

$$\bar{R}=E\{\min[Z(\omega)/D(s),1]\}$$

$$=\int_s^V D(v)/D(s)\,dR(v)+R(s).$$

The expected profit for the utility is $\bar{\pi}=(s-c)\bar{R}D(s)$, and for a consumer with willingness-to-pay v, the expected surplus is $(v-s)\bar{R}$.

With priority service, on the other hand, we can assign the probability of service $r(v)$ in such a fashion that for some value $\underline{v}$, $r(v)=R(v)$, if $v\geq\underline{v}$, and $r(v)=0$, if $v<\underline{v}$. (Notice that when $\underline{v}=c$, the resulting price menu is socially optimal.) We set the service charge equal to the constant s. Then the expected profit for the utility is

$$\pi(\underline{v})=-\int_0^V[p(v)+(s-c)r(v)]\,dD(v).$$

$$=-\int_{\underline{v}}^V[p(v)+(s-c)R(v)]\,dD(v).$$

For a consumer with willingness-to-pay v, the expected surplus is $(v-s)r(v)-p(v)$, and the incremental benefit from priority service is

$$B(v)=(v-s)\big[r(v)-\bar{R}\big]-p(v).$$

PROPOSITION 4: *If* $\bar{\pi}\geq\pi(c)$, *priority service is Pareto superior to random rationing with a fixed price.*

PROOF:

It follows from self-selection that B is nonnegative in the range $[0,s]$. Therefore, we only need to show that $B(v)\geq 0$ for $v\geq s$. The above formula implies that B is convex in this range, and $B'(s)\leq 0$ and $B'(V)\geq 0$; hence, its minimum occurs at an intermediate value v^* where $B'(v^*)=0$.

Recall that the priority service charge is

$$p(v)=p(s)+\int_s^v[r(v)-r(u)]\,du,$$

$$=p(s)+\int_s^v(u-s)\,dr(u),$$

from which it follows that $p'(v)=(v-s)r'(v)$ and therefore $r(v^*)=\bar{R}$ and $B(v^*)=-p(v^*)$. We must therefore show that $p(v^*)\leq 0$. The feature that $r(v^*)=\bar{R}$ reflects the fact that the least-advantaged customer is the one for whom the absence of priority service produces an optimal service probability.

Since $\pi(s)\geq\bar{\pi}$ and by the supposition $\bar{\pi}\geq\pi(c)$ (which implies $s\geq c$), there exists a value $\underline{v}$, which lies between c and s, such

that $\pi(\underline{v})=\bar{\pi}$. This implies that

$$p(s)D(s) = -\int_s^V (v-s)D(v)\,dR(v) + \int_{\underline{v}}^s [p(v)+(s-c)R(v)]\,dD(v).$$

In this expression, the second integral can be shown to be nonpositive as follows: By construction, $p(\underline{v})+sR(\underline{v})=\underline{v}R(\underline{v})$. It follows that $p(\underline{v})+(s-c)R(\underline{v})=(\underline{v}-c)R(\underline{v})\geq 0$. The above assertion follows from the property that p and R are nondecreasing and that D is nonincreasing. Therefore, we obtain

$$p(s)\leq -\int_s^V (v-s)D(v)/D(s)\,dR(v).$$

Using this result, and the property that $0=(v^*-s)[\bar{R}-R(v^*)]$, yields the two expressions

$$p(v^*)\leq -\int_s^V (v-s)D(v)/D(s)\,dR(v) + \int_s^{v^*}(v-s)\,dR(v),$$

$$0 = +\int_s^V (v^*-s)D(v)/D(s)\,dR(v) - \int_s^{v^*}(v^*-s)\,dR(v),$$

which when added produce

$$\begin{aligned} p(v^*) &\leq \int_s^V (v^*-v)D(v)/D(s)\,dR(v) - \int_s^{v^*}(v^*-v)\,dR(v) \\ &\leq \int_s^{v^*}(v^*-v)D(v)/D(s)\,dR(v) - \int_s^{v^*}(v^*-v)\,dR(v) \\ &= \int_s^{v^*}(v^*-v)[D(v)/D(s)-1]\,dR(v) \\ &\leq 0, \end{aligned}$$

since $D(v)\leq D(s)$, as required.

The implication of this result is that there exists a method of distributing part of the priority service revenues as dividends to customers that ensures that no customer's net benefit be reduced by the implementation of priority service. Distributional effects remain, and in particular those for whom existing service probabilities are optimal are the least advantaged, but no customer is worse off due to the adoption of priority service. One can plausibly expect unanimous support for a well-designed implementation of priority service.

V. Compensatory Insurance

A limitation of the basic model is the assumption that consumers are not averse to the risk of service curtailment. If some consumers are risk averse, then attainment of full efficiency requires that insurance be provided to compensate customers for the financial consequences of interruptions.

In principle, the following scheme suffices if a risk-neutral underwriter can write auxiliary insurance contracts for customers who elect them. According to the basic model, a consumer whose valuation of a unit of service is v is predicted to select the option $(p(v), s(v), r(v))$, where $p(v)$ is the priority service charge, $s(v)$ is the charge for delivered service, and $r(v)$ is the probability of service delivery. Consequently, a contract that pays the face amount $v-s$ in the event of interruption (nondelivery) fully insures the customer. The premium that compensates the underwriter for the actuarial risk of this contract is $(v-s)[1-r(v)]$. Offered a supplementary schedule of such contracts with their associated actuarial premia, a risk-averse customer whose valuation is v selects both the priority option $(p(v), s(v), r(v))$ and the supplementary insurance contract with the premium $(v-s)[1-r(v)]$ that pays $v-s$ if service is not delivered.

Alternatively, the utility can act as the agent for the mutual insurance association comprising the customers collectively. Assume that the consumers are sufficiently numerous that in the aggregate they share risks so finely that they can be considered

risk neutral. Taking the case that insurance coverage is fully bundled with the priority service, the utility offers a schedule consisting of options $(p, s, r; q, v)$ in which the insurance premium is $q = (v - s)[1 - r]$ and the compensation $v - s$ in the event of interruption is the predicted net valuation of a consumer selecting this option. The consumer pays $p + q$ initially, receives service at the price s with probability r, and with probability $1 - r$ is denied service but receives the compensation $v - s$. We express such an option in the reduced form $(P, s, r; v)$, where $P = p + q$.

Recall that efficient implementation of the basic model requires that the utility curtail service to customers in the order of their priority selections, since these reveal their valuations of service. That is, customers selecting options with particular values of p or r are not curtailed unless all customers selecting options with lower values are curtailed too. In the case of insurance, using the reduced form above, service is curtailed in the order of P, r, or v; in particular, we emphasize the latter possibility. The utility's rule is: select customers for interruption so as to minimize the compensation that must be paid. This rule also provides the correct incentives to an independent underwriter able to influence the selection of interrupted customers.

PROPOSITION 5: *A premium for priority insurance that is the sum of the actuarial risk and the priority service charge, and a service priority based on the insurance coverage selected, yields efficient risk sharing and efficient rationing.*

PROOF:

A consumer with valuation v selecting the option $(P(\hat{v}), s; \hat{v})$ acquires the lottery that pays $v - s - P(\hat{v})$ and $\hat{v} - s - P(\hat{v})$ with some probabilities $\hat{r}$ and $1 - \hat{r}$. Efficient risk sharing (with a risk-neutral underwriter) requires that the consumer select $\hat{v} = v$, so that he is fully covered. A risk-averse consumer will do this only if $P'(v) = 1 - \hat{r}$, as one can verify. Efficient rationing of scarce supplies requires that $\hat{r} = R(v)$, which is the probability that the spot price does not exceed v; therefore, $P'(v) = 1 - R(v)$. Thus, the premium charged for each \$1 increment in coverage should be the actuarial value of the risk that this increment will be paid. Integrating this differential equation subject to the boundary condition $P(s) = 0$ yields $P(v) = [v - s][1 - R(v)] + p(v)$, where $p(v)$ is the priority service charge calculated according to Theorem 1. Thus, the total premium is precisely the actuarial premium for the compensatory insurance, plus the priority service charge.

Since the insurance premium for full coverage at actuarially fair rates can always be separated in this fashion, in the following we ignore the effects of risk aversion and the role of insurance. We mention that if compensatory insurance is offered for interruptions due to capacity limitations, it can also be augmented to insure against interruptions due to transmission failures.

VI. An Example

We illustrate some of the previous results with an example. Suppose that there is a single-generation technology with marginal cost $c = 0$ that has probability $F(q) = [q/K]^a$ of providing a supply no more than q, and the demand function is $D(v) = [1 - v]^b$. One can interpret $K \geq 1$ as the potential capacity of the system. Note that the maximum demand is $D(0) = 1$ and the maximum valuation is $V = 1$. With random rationing of scarce supplies, the probability that any unit is served is

$$\overline{R} = 1 - \left[1 - \frac{ab}{ab + b}\right] / K^a,$$

while with efficient rationing the probability that a unit with the valuation v is served is

$$R(v) = 1 - [1 - v]^{ab} / K^a.$$

The menu that induces self-selection by each customer imposes a priority service

charge of

$$p(v)=p(0)+\frac{ab}{K^a}\left\{\frac{(1-v)^{ab+1}-1}{ab+1}-\frac{(1-v)^{ab}-1}{ab}\right\},$$

and if the revenues are redistributed to customers as dividends then the net effect is that

$$p(0)=-\frac{ab}{K^a}\left\{\frac{1}{ab+b}-\frac{1}{ab+b+1}\right\}.$$

The customer least advantaged by the adoption of priority service has valuation

$$v^*=1-\left[1-\frac{ab}{ab+b}\right]^{1/ab}.$$

The complexity of the formulas escalates from this point; so to be more specific we consider the special case that $a=b=K=1$. With these parameters, $p(v)=p(0)+v^2/2$ and $p(0)=-1/6$. Also, $\bar{R}=1/2$ and $v^*=1/2$. If insurance provisions are bundled with priority service, then the premium for coverage in the amount v is $P(v)=1/3-[1-v]^2/2$ if expected net revenues are to be zero. Alternatively, if priority points are used (see Section IX), then x points assigned to a unit gains the customer a service reliability of $\hat{r}(x)=\sqrt{2x}$ if $x\le 1/2$ (this excludes any dividends). Including dividends, the net benefit from priority service for a customer with the valuation v is

$$B(v)=\tfrac{1}{24}+\tfrac{1}{2}\left[v-\tfrac{1}{2}\right]^2.$$

The minimal benefit $B(v^*)$ is only 25 percent of the maximal benefit obtained by customers with the two extreme valuations.

VII. Finite Number of Priority Classes

When transaction costs are recognized, it is expected that there will be only a finite number of priority classes. In this section we show that the incremental gains from priority service decline rapidly as the number of priority classes increases. Therefore, a few priority classes can capture most of the potential benefits from priority service. To see this, we continue to use the previous example with a linear demand function and a uniform distribution of supply. Consumers are divided into n priority classes based on their willingness to pay, say, $[0,v_1]$, $[v_1,v_2],\cdots[v_{n-1},1]$, where $0=v_0<v_1<\cdots<v_{n-1}<v_n=1$. Suppose that the service is provided to the consumers in such a manner that consumers in a higher value class are given a higher priority (and pay more), but within each class, all consumers are treated equally and therefore are served in a random order. Then, the probability that a consumer with valuation v between v_i and v_{i+1} will be served is

$$\begin{aligned}r(v)=r_i&=\int_{v_i}^{v_{i+1}}\left[\frac{D(v)-D(v_{i+1})}{D(v_i)-D(v_{i+1})}\right]dR(v)\\&\quad+R(v_i)\\&=(v_{i+1}+v_i)/2,\end{aligned}$$

and

$$p(v)=p_i=v_0r_0+\sum_{j=1}^{i}v_j(r_j-r_{j-1}).$$

The expected social surplus is

$$\begin{aligned}S_n&=\left|\int_0^V vr(v)\,dD(v)\right|\\&=\sum_{i=0}^{n-1}\int_{v_i}^{v_{i+1}}vr(v)\,dv\\&=\sum_{i=0}^{n-1}(v_{i+1}+v_i)(v_{i+1}^2-v_i^2)/4.\end{aligned}$$

Optimal market segmentation requires that

$$\begin{aligned}\partial S_n/\partial v_i&=0\\&=(v_{i+1}-v_i)^2-(v_i-v_{i-1})^2,\end{aligned}$$

for every i.

It follows immediately that

$$v_{i+1} - v_i = v_i - v_{i-1},$$

and thus

$$v_i = i/n, \quad \text{for every } i.$$

Using this result, we can write

$$S_n = \sum_{i=0}^{n-1} (2i+1)^2/4n^3 = \frac{1}{3}\left(1 - \frac{1}{4n^2}\right).$$

This suggests that the unrealized surplus ($S_\infty - S_n = 1/12n^2$) declines rapidly as the number of priority classes increases. In this example, priority service increases the total social surplus by 33 percent. Of this amount, nearly 90 percent can be captured by offering just three priority classes.

The following proposition shows that the properties of this example are quite generally true asymptotically if demand is nonstochastic. Let P be the inverse demand function, namely $D(P(q)) = q$, and assume that P and the supply distribution F have bounded second derivatives on the interval $[0, Q]$, where $P(Q) = 0$. Continue to assume that the marginal cost of supply is constant and normalized to be zero.

PROPOSITION 6: *The surplus that is unrealized due to a finite number n of priority classes is of order $1/n^2$. That is, $S_n \geq S_\infty - 0(1/n^2)$.*

PROOF:

Define $G(q) = 1 - F(q)$ so that the potential surplus is

$$S_\infty = \int_0^Q P(q)G(q)\,dq.$$

For the optimal market segmentation with n priority classes, let $q_i = D(v_i)$, where $q_n = 0 < \ldots < q_0 \leq Q$, so that $\Delta_i = q_{i-1} - q_i$ is the number of customers served by the ith priority class. The realized surplus can then be expressed as

$$S_n = \sum_{i=1}^{n} \bar{v}_i r_i \Delta_i,$$

where for class i the average valuation and service reliability are

$$\bar{v}_i = \int_{q_i}^{q_i+\Delta_i} P(q)\,dq/\Delta_i,$$

$$r_i = \int_{q_i}^{q_i+\Delta_i} G(q)\,dq/\Delta_i,$$

respectively. According to the Hermite interpolation formula (Anthony Ralston, 1965, Sec. 4.5):

$$\bar{v}_i = v_i + \tfrac{1}{2}\Delta_i P'(q_i) + 0(\Delta_i^2),$$

$$r_i = g_i + \tfrac{1}{2}\Delta_i G'(q_i) + 0(\Delta_i^2),$$

where $g_i = G(q_i)$. Consequently,

$$S_n = \sum_{i=1}^{n} \left[\bar{v}_i + \tfrac{1}{2}\Delta_i P'(q_i)\right] \times \left[g_i + \tfrac{1}{2}\Delta_i G'(q_i)\right]\Delta_i + 0(\Delta_i^3),$$

$$= \sum_{i=1}^{n} \bar{v}_i g_i \Delta_i + \tfrac{1}{2}\left[P(q_i)G(q_i)\right]'\Delta_i^2 + 0(\Delta_i^3).$$

Using the fact that the boundaries of the priority classes are chosen to maximize the realized surplus S_n, we know that $S_n \geq S_n^*$, where S_n^* is the realized surplus if the priority classes are restricted to be all the same size; namely, $q_i = Q(1 - i/n)$ and $\Delta_i = \Delta = Q/n$. Since the formulas above apply also to S_n^*, we have

$$S_n \geq S_n^* = \left\{\sum_{i=1}^{n} \bar{v}_i g_i \Delta\right\} + \tfrac{1}{2}\Delta\left\{\sum_{i=1}^{n} \left[P(q_i)G(q_i)\right]'\Delta\right\} + 0(\Delta^2).$$

According to the "trapezoid rule" for numerical integration (Ralston, 1965, Eq.

(4.12-11)), the bracketed expressions are

$$\sum_{i=1}^{n} \bar{v}_i g_i \Delta = \int_0^Q P(q)G(q)\,dq + \tfrac{1}{2}\Delta[v_0 g_0 + v_n g_n] - 0(\Delta^2),$$
$$= S_\infty + \tfrac{1}{2}\Delta[v_o g_o + v_n g_n] - 0(\Delta^2);$$
$$\sum_{i=1}^{n} [P(q_i)G(q_i)]'\Delta = \int_0^Q [P(q)G(q)]'dq + 0(\Delta),$$
$$= [P(Q)G(Q) - P(0)G(0)] + 0(\Delta).$$

Now $v_0 = P(q_0) = P(Q) = 0$ and $v_n g_n = P(0)G(0)$, and therefore

$$S_n \geq S_n^* = S_\infty - 0(\Delta^2),$$

which is the result to be proved.

An important implication of this result is that a few priority classes suffice to realize most of the efficiency gains from priority service. In particular, if transaction costs are significant, then priority service with several classes offered can be superior to spot markets.

VIII. Optimal Investment

We now turn to the problem of optimal investment policy. We begin by writing the social welfare function as follows:

$$(24)\quad E\left\{\int_0^\infty [P(z,\omega) - C(z,\omega)]^+ dz\right\} - \sum_{j=1}^{n} k_j X_j$$
$$= \sum_{j=1}^{n} E\left\{\int_{Z_{j-1}(\omega)}^{Z_j(\omega)} [P(z,\omega) - c_j]^+ dz\right\} - \sum_{j=1}^{n} k_j X_j.$$

Differentiating (24) with respect to X_i, for $i = 1,\dots,n$, yields the following optimality conditions

$$(25)\quad \sum_{j=1}^{n} E\{[P(Z_j(\omega),\omega) - c_j]^+ \,\partial Z_j(\omega)/\partial X_i\} - \sum_{j=1}^{n} E\{[P(Z_{j-1}(\omega),\omega) - c_j]^+ \,\partial Z_{j-1}(\omega)/\partial X_i\} = k_i,$$
$$\text{for } i = 1,\dots,n.$$

In view of (10), we have

$$(26)\quad \partial Z_j(\omega)/\partial X_i = \begin{cases} 0, & \text{if } j < i; \\ \partial Y_i(X_i,\omega)/\partial X_i, & \text{if } j \geq i. \end{cases}$$

Recall that the unit availability factor is $a_i = E\{\partial Y_i(X_i,\omega)/\partial X_i\}$. Substituting (26) in (25) and using the assumption of independent capacity increments yields

$$(27)\quad \sum_{j=1}^{n} E\{[P(Z_j(\omega),\omega) - c_j]^+\} a_i - \sum_{j=i+1}^{n} E\{[P(Z_{j-1}(\omega),\omega) - c_j]^+\} a_i = k_i,$$
$$\text{for } i = 1,\dots,n-1,$$

and

$$(28)\quad E\{[P(Z_n(\omega),\omega) - c_n]^+\} a_n = k_n.$$

Substituting (28) in (27) and repeating such a substitution backward yields

$$(29)\quad E\{[P(Z_i(\omega),\omega) - c_i]^+\} - E\{[P(Z_i(\omega),\omega) - c_{i+1}]^+\} = (k_i/a_i) - (k_{i+1}/a_{i+1}),$$
$$\text{for } i = 1,\dots,n-1.$$

In view of Lemma 1, we can rewrite (28) and (29) in the following form,

$$E\{[\hat{p}(\omega)-c_n]^+\}=k_n/a_n \tag{30}$$

and

$$E\{[\hat{p}(\omega)-c_i]^+\} - E\{[\hat{p}(\omega)-c_{i+1}]^+\} = (k_i/a_i)-(k_{i+1}/a_{i+1}), \quad \text{for } i=1,\dots,n-1. \tag{31}$$

Intuitively, the left-hand side of (30) can be interpreted as the expected profits resulting from adding an effective unit of technology n, and the right-hand side represents the capacity cost required for such an expansion. The optimal capacity level is achieved when the expected profits from an additional unit of technology n exactly cover the incremental capacity cost.

Similarly, the left-hand side of (31) can be interpreted as the expected increase in profits resulting from substituting an effective unit of technology $i+1$ by one of technology i, and correspondingly, the right-hand side represents the capacity cost increase due to this exchange. Therefore, the optimal capacity mix is achieved when the expected increase of profits equals the increased capacity cost resulting from such an exchange.

The optimality conditions (30) and (31) can be simply expressed in terms of priority charges.

PROPOSITION 7: *The optimal capacity configuration satisfies the following conditions*:

$$k_i/a_i+c_i+p^*(c_i)=k_1/a_1+c_1, \quad \text{for } i=1,\dots,n. \tag{32}$$

$$p^*(V)=p^*(c_n)+k_n/a_n. \tag{33}$$

PROOF:

First, the left-hand side of (30) can be written

$$\begin{aligned} E\{[\hat{p}(\omega)-c_n]^+\} &= \int_0^V \Pr\{p(\hat{\omega})-c_n>u\}\,du \\ &= \int_{c_n}^V \Pr\{\hat{p}(\omega)>u\}\,du \\ &= \int_{c_n}^V [1-r^*(u)]\,du \\ &= p^*(V)-p^*(c_n). \end{aligned} \tag{34}$$

(From equations (15) and (16).)

The expression (33) follows from (34) and (30).

Similarly, the left-hand side of (31) can be written

$$\begin{aligned} E\{[\hat{p}(\omega)-c_i]^+\} &- E\{[\hat{p}(\omega)-c_{i+1}]^+\} \\ &= \int_0^V \Pr\{\hat{p}(\omega)-c_i>u\}\,du - \int_0^V \Pr\{\hat{p}(\omega)-c_{i+1}>u\}\,du \\ &= \int_{c_i}^V \Pr\{\hat{p}(\omega)>u\}\,du - \int_{c_{i+1}}^V \Pr\{\hat{p}(\omega)>u\}\,du \\ &= \int_{c_i}^{c_{i+1}} \Pr\{\hat{p}(\omega)>u\}\,du \\ &= \int_{c_i}^{c_{i+1}} [1-r^*(u)]\,du \\ &= c_{i+1}-c_i+p^*(c_{i+1})-p^*(c_i). \end{aligned} \tag{35}$$

(From equations (15) and (16).)

The expression (32) follows from (35) and (31).

The above formulas establish a simple relationship between priority charges and the capacity and operating costs in the long run. A key advantage of priority service is that by allowing consumers to self-select, it provides information about consumers' willingness to pay for capacity increments. Otherwise, the optimal capacity must depend on consumers' valuation of average service reliability, which cannot be measured directly, and is hard to estimate accurately (Michael Telson, 1975, and Hung-po Chao, 1983).

IX. Alternative Forms of Market Organization

The implementation of priority service could take several organizational forms, which have considerably different implications for costs and risks. In this section we describe three implementations that differ in terms of the contract forms and the market organization.

In practice, consumers purchase multiple units, rather than a single unit (say, one kilowatt of power). A consumer's demand comprises several units with different valuations. These units can be ordered by their valuations, from a base unit with a high valuation down to a marginal or peak unit with a lower valuation. Recognition of this feature suggests the practicality of offering priority service in terms of "demand subscription," as it is known in the electric power industry. This form of implementation allows each consumer to select different reliability levels and corresponding rates for different increments of his demand. In such a scheme, the responsibility for estimating the chances of interruption and for interpreting contractual obligations rests primarily with the utility. However, this poses practical difficulties. First, the utility usually has imperfect knowledge of the distribution of consumers' valuations. A misspecification of the price menu could result in too few customers selecting low priority to enable provision of the higher level of service required for high-priority customers. Further, ambiguities in the interpretation and enforcement of such a contract may arise, unless the contracts are specified in terms of observable events.

In the second form of implementation, consumers purchase service insurance and can expect to be compensated for an interruption by an amount that depends on the insurance premium paid in advance. Then during supply shortages, the utility will first interrupt the service of those consumers who selected the lowest coverage. In such a scheme, the utility is committed to the priority ranking determined by the risk premium or interruption compensation stipulated unambiguously in the service contract, but not to a probability or frequency of service, though these may have been used to design the menu and to inform customers about the predicted consequences of their selections. Michael Manove (1983) shows that, in general, if the insurance is provided by the producer, this scheme will be free from the moral hazard problem, and it will induce both the producer and consumers to reduce losses efficiently. Therefore, this form of implementation requires relatively little monitoring and control.

In the third implementation form, each consumer buys priority points, which are then assigned to demand segments. A market will be created to allow consumers to exchange their priority point holdings. The utility is relieved of the task of designing a price menu. In an emergency, the utility curtails those demand units assigned the fewest priority points. In this approach the burden of assessing the likelihood of interruption is transferred to the market maker and participants. The market transactions of priority points will provide relevant information about the distribution of consumer valuations and a direct indication to the utility of whether capacity expansion is justified.

However, an efficient implementation of this scheme requires that customers' expectations be "rational," in the sense that their selections are based on reliability assessments that are eventually consistent and correct. Suppose that the utility offers to sell, for a unit price each, as many points as are demanded by customers, or to buy back at this price any unused points. The offer specifies a service charge $\hat{s}(x)$ possibly depending on the number x of points assigned. The customers' selections are based on an assessment that the probability of receiving

service is $\hat{r}(x)$. Rational expectation implies that $\hat{r}(x(v)) = R(v) = \Pr\{\hat{p}(\omega) \leq v\}$, or equivalently $\hat{r}(x) = \Pr\{p(\hat{p}(\omega)) \leq x\}$, where $\hat{p}(\omega)$ is the state-contingent spot price. Given this assessment, a customer with valuation v buys $x(v)$ points so that his net surplus, $(v - \hat{s}(x))\hat{r}(x) - x$, is maximized. This mechanism is formally equivalent to an implementation of priority service with the menu $M = \{(x, \hat{s}(x), \hat{r}(x))\}$. The question is under what condition this menu will induce an efficient allocation. As shown previously, a menu is efficient if $r(v) = R(v)$, as is implied by the rational expectation assumption. These arguments can be summarized as follows.

PROPOSITION 8: *The priority point scheme achieves an efficient rationing of scarce supplies, if consumers' selections reflect rational expectations.*

Two amendments can add to the merits of this implementation. First, the utility can systematically vary the price for points to account for daily and seasonal variations in demand and supply conditions; thus, customers need not change their point assignments unless their circumstances change relative to these cyclical variations. This is especially useful if consumers can change their point selections at any time. Alternatively, the utility can auction to brokers a limited supply of points and then let the price vary according to market conditions. This version has the advantage that it provides incentives to brokers and others to assess accurately the relevant probability distributions, and it enables a market in futures contracts.

X. Conclusion

The main feature of priority service is the product differentiation that it achieves. The uniform quality imposed by random rationing is replaced by a spectrum of qualities, each priced to induce an efficient selection by customers according to their valuations of service. The efficiency gains from this increased variety are derived directly from the increased efficiency of the resulting rationing of scarce supplies, and in the long run from the efficient provision of generating capacity. Under very weak conditions, priority service is Pareto superior to random rationing with a fixed price: every customer's net benefits are increased without reducing the seller's net revenue. The incremental gains from priority service decline quickly, however, as the number of priority classes is increased; thus, a few priority classes suffice to capture most of the efficiency gains.

Under the priority service scheme, perfect rationing is feasible through an appropriately designed price menu, although the individual consumers' preferences are only privately known. In such a menu, each priority option is so priced that the expected total charge exactly compensates for the losses incurred by those who choose lower priorities in order to provide the service reliability required for the higher priorities. In the long run, when capacity planning is included, the priority charge reflects the related capacity and operating costs. Thus, priority service provides the key information about customers' willingness to pay for capacity increments that previously was unavailable from any observed choice behavior of customers.

We have also compared the welfare implications of priority service and spot pricing. In the absence of transaction costs, these two schemes are equally efficient. And, in the absence of risk aversion, both are equivalent from the perspective of individual consumers. From the producer's perspective, however, they are not necessarily equivalent, unless the marginal demand function is stochastically independent of the aggregate demand and supply conditions. When the role of customers' risk aversion is recognized, efficient risk allocation requires that compensatory insurance be provided and that the risk neutral producer bear all the risk.

Priority service can be implemented in a variety of forms. We have mentioned auxiliary compensatory insurance to deal with customers' risk aversion and a market for priority points to implement the scheme in a simple workable fashion. The rationing order can, in principle, be based on any one of several indices: the customer's selected probability of service, the rank order of this selection, the priority service charge (particu-

larly in the case of priority points), or the insurance premium or compensation (in the case of bundled insurance provisions).

REFERENCES

Caramanis, M., Bohn, R. E. and Schweppe, F. C., "Optimal Spot Pricing: Practices and Theory," *IEEE Transactions on Power Apparatus and Systems*, *PAS-101*, 1982, 3234–45.

Chao, Hung-po, "Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty," *Bell Journal of Economics*, Spring 1983, *14*, 179–90.

______, **Oren, S., Smith, S. and Wilson, R.,** "Multilevel Demand Subscription Pricing for Electric Power," *Energy Economics*, October 1986, *4*, 199–217.

Harris, M. and Raviv, A., "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *American Economic Review*, *71*, June 1981, 347–65.

Manove, Michael, "Provider Insurance," *Bell Journal of Economics*, Autumn 1983, *14*, 489–96.

Marchand, M. G., "Pricing Power Supplied on an Interruptible Basis," *European Economic Review*, October 1974, *5*, 263–74.

Oren, S., Smith, S., Wilson, R. and Chao, H., "Priority Service: Unbundling the Quality Attributes of Electric Power," EPRI Report, EA-4851, Electric Power Research Institute, Palo Alto, CA, 1986.

Ralston, Anthony, *A First Course in Numerical Analysis*, New York: McGraw-Hill, 1965.

Telson, M. L., "The Economics of Alternative Levels of Reliability for Electric Power Generation Systems," *Bell Journal of Economics*, Autumn 1975, *6*, 679–94.

Tschirhart, J. and Jen, F., "Behavior of Monopoly Offering Interruptible Service," *Bell Journal of Economics*, Spring 1979, *10*, 244–58.

Vickrey, William, "Responsive Pricing of Public Utility Services," *Bell Journal of Economics and Management Science*, 1971, *2*, 337–46.

The Relative Rigidity of Monopoly Pricing

By JULIO J. ROTEMBERG AND GARTH SALONER*

This paper examines why monopolies change their nominal prices less often than do tight oligopolies. We show that cost (demand) changes create a larger incentive for duopolists (monopolists) to change their prices. When both costs and demand are affected by small changes in the overall price level, the cost effect dominates. In the presence of a small, fixed cost of changing prices, therefore, duopolists change their prices in response to smaller perturbations in underlying conditions.

The relationship between industry structure and pricing is a major focus of industrial organization. One of the most striking facts to have emerged about this relationship is that monopolies tend to change their prices less frequently than tight oligopolies. Although the first evidence in this regard was presented by George Stigler (1947) almost 40 years ago, no theoretical explanations have been offered. The objective of this paper is to develop a model capable of explaining these facts.

Stigler's objective in comparing the relative rigidity of monopoly and duopoly prices was to test the kinked demand curve theory of R.L. Hall and C.J. Hitch (1939) and Paul Sweezy (1939). Since the work of Gardiner Means (1935) seemed to show that concentrated industries exhibited greater price rigidity than their unconcentrated counterparts, the kinked demand curve was developed and embraced as providing a theoretical foundation for the rigidity of prices. It was widely regarded to be an implication of that theory that duopolists would not change their prices in response to small changes in their costs.[1]

Stigler's test (1947) was a direct and simple test of the rigidity of oligopoly prices. Instead of comparing oligopoly pricing with pricing in unconcentrated industries, he simply compared the relative rigidity of monopoly and oligopoly prices. If it is the kink that leads to inflexible oligopoly prices, monopolists should have more flexible prices since monopolists do not face a kinked demand curve. Stigler found instead that monopolist's prices were even more rigid. Several later empirical studies have supported his original finding: monopolists change their prices less frequently than do oligopolists. This finding throws into question any theory in which prices are rigid only because individual oligopolists fear "upsetting the applecart."

In his study, Stigler tabulated the number of price changes for two monopolistically supplied commodities (aluminum and nickel) and 19 products, which were each supplied by a small number of firms. The source of the data on price changes was the Bureau of Labor Statistics (BLS) bulletins, *Wholesale Prices,* for the period June 1929–May 1937. The price of nickel did not change at all over this period and there were only two price changes for aluminum. Among the oligopolistically supplied products, however, only one had fewer than four price changes (sulfur) and half had more than 10 price changes.

*Sloan School of Management and Department of Economics, MIT, Cambridge, MA, 02139, respectively. We would like to thank Olivier Blanchard and Gregory Mankiw for helpful discussions, two anonymous referees for their suggestions, and the National Science Foundation, grant nos. SES-8209266 and IST85-10162, respectively, and the Hoover Institution for financial support.

[1]This view was emphasized by Martin Bronfenbrenner (1940), for example. It is now well recognized, however, that the kinked demand curve implies multiple equilibria. When cost conditions change, one might well expect the equilibrium to change as well. It is only if the current price is somehow "focal" that the price will not change.

The finding has been replicated on other price data for different periods. Julian Simon (1969) studied advertising rates of business magazines from 1955 to 1964. Simon's data have the advantage that they contain the price series for each publication rather than simply an index of the industry price, as is the case for the BLS data. Simon finds that the larger the number of competitors a firm has, the higher the number of years in which it changes its price.

Walter Primeaux and Mark Bomball (1974) compare pricing of electric power when it is supplied by a duopoly versus a monopoly.[2] Their data cover 17 duopolies and 22 monopolies from 1959 to 1970. One advantage of their study is that there is no product differentiation in electric power, so that there is no danger of misspecifying the firms' competitors. Also, list prices are transactions prices since deviations from printed schedules are illegal for public utilities. They show that when there are two firms in the market they each change their prices more often than a corresponding monopolist. This is true for all levels of power usage. The effect is more pronounced for lower levels of power usage, where the duopolists changed their prices two or three times more frequently than it is for higher-power usages where they changed their prices roughly one-and-a-half times as often.

Finally, Primeaux and Mickey Smith (1976) study the pricing of 88 major drugs during the period 1963–73. For their sample they are unable to reject the hypothesis that oligopolists and monopolists have the same frequency of price changes.

These studies suggest that monopolists generally change their prices less frequently than oligopolists. The explanation that we offer for this phenomenon is based on the relative incentives of monopolists and oligopolists to adjust their prices when underlying cost and demand conditions change or when inflation erodes existing prices. We focus on the comparison between duopoly and monopoly and show that whether the products of the duopoly are homogeneous or differentiated, the duopolists have a greater incentive to change their prices than does a monopolist, facing the same configuration of demand. So, if there are other forces leading to price rigidity that are of roughly comparable magnitudes across industry structures, there will be a general tendency for duopoly prices to change more frequently. The particular reason why prices may be unresponsive to changes in underlying conditions that we focus on is that there may be a fixed cost of changing prices. This idea was first put forward by Robert Barro (1972) and has been used by Eytan Sheshinski and Yoram Weiss (1977, 1985), Julio Rotemberg (1983), and Gregory Mankiw (1985), among others. These costs are usually taken to include the physical costs of changing and disseminating price lists and the possibility of upsetting customers with frequent price changes. In the electric utility industry they also capture the costs of obtaining permission from regulatory authorities to change tariffs.

If it is costly for firms to change their prices, the question then is how the gains to the firms of changing their prices compare with the costs and, more importantly, how the gains differ across market structures. To see why duopolists in general have a greater incentive to change prices in response to costs changes, consider the following simple case. Suppose that two firms competing in Bertrand style and charging price equal to constant marginal costs unexpectedly discover that costs have increased. If neither firm increases its price, the firms share the loss of supplying the entire market demand at a price below costs. Each firm obviously has a large incentive to change its prices. Furthermore, if either firm believes its rival will change its price, then it has an even greater incentive to raise its own price in order to avoid suffering the entire loss itself. Put differently, when a firm changes its own price

[2] The duopoly cases were situations in which a municipally owned electric utility and a privately owned firm competed directly. Each firm had its own electricity generation facility, and customers had the choice of which firm they wished to be served by. In some cases, customers could switch from one company to the other at will; in other cases, new customers had a choice of supplier but once they had made a choice, it was not possible to switch suppliers.

it imposes a negative externality on its rival: it increases the amount that the firm must sell at the "wrong" price.

A similar phenomenon arises for cost decreases. In that case there is no incentive for the firms to make a combined price decrease. However, there are substantial incentives for either firm to make a unilateral price decrease to undercut the rival. Here again there is an externality: the deviating firm's gain is made at the rival's expense.

A monopolist's profits are differentiable in its price. Therefore, as George Akerlof and Janet Yellen (1985) show, the loss in profits from not changing its price is second order. Since the duopolist's incentives to change price are first order, if they face comparable costs of changing prices, the duopolists would change price more frequently in response to a cost change than a monopolist would.

The reverse is true for changes in demand. Since Bertrand competitors set price equal to marginal cost, they have no incentive to change price in response to a shift in demand. A monopolist, of course, does have an incentive to change price in these circumstances.

When oligopolists produce differentiated products, individual firms' profits are again differentiable in their own prices and the Akerlof-Yellen argument still applies. One might believe, therefore, that the result that monopolists change their prices less frequently than duopolists in response to a cost change would not hold true in the case of differentiated products. In fact it does. The reason has to do with the externalities discussed above.

Consider duopolists producing differentiated products and, as above, suppose that costs increase slightly. Now if one firm raises its price slightly, it no longer yields all of its customers to its rival. Profits are no longer discontinuous at the point of equal prices. However, it does lose some of its customers to its rival, and if the degree of substitutability is high, it loses them at a rapid rate. In other words, the externality that the duopolist inflicts on its rival is increasing in the degree of substitutability between the products. Thus, the increase in profits from adjusting its price may be large. A monopolist, on the other hand, is able to internalize these externalities. For purposes of comparison, suppose that the monopolist offers both products. Now when it changes the price of either product it bears the full consequences: both the change in profits of the product whose price is changed and that of the product whose price is unchanged. Whereas the duopolists each have an incentive to change price in order to make a gain at the other's expense, the monopolist has no such incentive.

Thus even in the case of differentiated products, provided the degree of substitutability between the products is great enough, duopolists have a greater incentive to change their prices in response to a change in costs than a monopolist does. The incentive to change price in response to a change in demand, however, is greater for the monopolist. In order to see which effect is likely to dominate in practice, we examine the situation when both costs and demand are affected by overall changes in the price level. We find that for small changes in the price level, the cost effect dominates the demand effect and so duopolists have a greater incentive to change their prices than a monopolist does. Thus in the presence of fixed costs of changing prices the monopolist may adjust prices more sluggishly.

In Section I we develop intuition for our result via a homogeneous goods example. This model is generalized to differentiated products and an inflationary environment in Section II. We conclude with Section III.

I. A Model with Homogeneous Products

Throughout the paper we assume that duopolists treat prices as their strategic variables. This seems natural given that we focus on whether firms change their prices or not. In this section, in order to demonstrate how the incentives for a monopoly to change prices differ from those of a duopoly, we begin with a very simple model. In particular, we will assume that the duopolists produce a homogeneous good with constant (and equal) marginal costs. As we shall see below,

this formulation is useful for expository purposes since the incentives for changing prices are most apparent when the model is stripped down in this way.

Unfortunately, we will also see that this formulation is too stark in the sense that duopolists earn zero-profits gross of any fixed costs. Thus, if they must bear any such costs, their participation becomes unprofitable. However, any number of modifications in the direction of realism (such as differentiated products or increasing marginal costs) would provide the firms with sufficient profits to cover small fixed costs. We begin with the simplest model to develop the intuition for the result. We later consider product differentiation where the existence of profits guarantees the willingness of the firms to participate as long as costs of changing prices are small.

Time is divided into two "periods" by an unexpected increase in the firms' constant marginal costs of production from c_1 to c_2. We will refer to the periods before and after the cost change as periods 1 and 2, respectively. Industry demand is given by $q = a - bP$, $a/b > c_2$, where P is the lowest price charged. Since the duopolists compete in Bertrand style $P_1^1 = P_1^2 = c_1$ (subscripts denote periods and the firm is indexed by the superscript). The monopolist, on the other hand, charges $P_1^m = (a + bc_1)/2b$ and sells $(a - bc_1)/2$.

We explore how the change in costs affects prices. We consider what happens when the new level of marginal costs, c_2, is known to both firms before they select their period 2 prices, but where each firm must incur a a fixed cost, f, to change its price.

If the monopolist leaves its price unchanged at P_1^m, it sells $(a - bc_1)/2$ and earns $[\{(a + bc_1)/2b\} - c_2]\{(a - bc_1)/2\}$. If, on the other hand, it changes its price, it earns $(a - bc_2)^2/4b - f$. It is therefore worthwhile for it to change its price if and only if

$$(1) \qquad \frac{b(c_2 - c_1)^2}{4} > f.$$

Now consider a duopolist. The amount demanded at $P = c_1$ is $q_1 = a - bc_1$. Suppose that firm 2 does not change its price. If firm 1 does not change its price either, the firms share the loss of $q_1(c_2 - c_1)$, that is, they each lose $(a - bc_1)(c_2 - c_1)/2$. What happens if firm 1 increases its price? To do so it must incur the cost, f. It then loses all its sales to firm 2. Thus firm 1 loses f if it raises its own price and firm 2 keeps its price unchanged. So, in this case, firm 1 prefers to change its price if

$$(2) \qquad (a - bc_1)(c_2 - c_1)/2 > f.$$

Now consider what happens if firm 2 increases its price to c_2. Now firm 1 loses $(a - bc_1)(c_2 - c_1)$ if it maintains its period 1 price (since it now bears the entire loss itself). On the other hand, it loses only f if it joins firm 2 in the price increase to c_2. Thus it prefers to raise its price if

$$(3) \qquad (a - bc_1)(c_2 - c_1) > f.$$

Equation (2) implies equation (3). Thus if (2) holds, changing price is a dominant strategy and the unique equilibrium involves both firms changing price. If (3) holds but (2) does not, each firm is willing to change its price only if the other also does. There are then two pure strategy equilibria: one in which the firms both change their prices and one in which neither does. Finally, if (3) does not hold, then the unique equilibrium is that neither firm changes its price.

Now compare the relative incentives for the duopoly and the monopoly to change prices. To make the comparison unfavorable to frequent price changes by the duopoly, we concentrate on the case in which changing price is the unique equilibrium. Then the duopoly changes prices if (2) holds while the monopoly changes prices if (1) holds. Since $a/b > (c_1 + c_2)/2$ by assumption, if (1) holds then (2) holds as well. Thus the duopolists would always change the price if the monopolist would. Moreover, if $(a - bc_1) > 2f/(c_2 - c_1) > b(c_2 - c_1)/2$, then (2) holds but (1) does not, so that, for parameters in this range, the duopolists would change their prices whereas the monopolist would not.

The intuition for these results is clear from Figure 1, which illustrates the effect of a cost increase. The profit for a monopolist who sets the optimal price for costs c_2 is given by the integral of marginal revenue minus marginal costs evaluated at q_2^m. This is equal to the shaded area in Figure 1. If the monopolist does not change its price (so that it sells q_1^m), it earns the profits it would earn if its costs were actually c_1 (the area ac_1z) minus $(c_2-c_1)\ q_1^m = c_1c_2yz$. The loss from not changing its price is therefore the crosshatched triangle

$$(4) \qquad xyz = (c_2-c_1)(q_1^m-q_2^m)/2$$
$$= b(c_2-c_1)^2/4.$$

The monopolist is willing to change its price if this area exceeds f.

Now consider the duopolists. If firm 1 believes that firm 2 will not change its price, firm 1 can raise its price to c_2 and earn zero (less the fixed cost f). On the other hand, if it does not change its price, it shares the industry loss of c_1c_2vw. Clearly, $(c_1c_2vw)/2$ always exceeds (xyz). Thus the duopolist always has a greater incentive to increase its price.[3]

From Figure 1 we can see that the greater incentive for the duopolist to change its price does not hinge on the linearity of demand. For an arbitrary demand function and its corresponding marginal revenue function, the relevant question is how the area corresponding to the crosshatched area xyz compares with the corresponding $(c_2c_1wv)/2$.

Notice that $(q_1^m-q_2^m)(c_2-c_1)$ is an upper bound for the area xyz and the area $(c_2c_1wv)/2$ is equal to $q_1^c(c_2-c_1)/2$, where q_1^c is the duopoly output in period 1. Therefore the area xyz is less than $(c_2c_1wv)/2$ if $q_1^m-q_2^m < q_1^c/2$, that is, if $q_1^c > 2(q_1^m-q_2^m)$.

[3]This result should not be taken to mean that duopolists have a larger incentive to respond to changes in costs whatever the form of their strategic interaction. For instance, if Cournot duopolists with linear demand and constant marginal costs face an increase in these costs, each firm's loss from not adjusting its quantity exactly equals the corresponding loss for a monopolist.

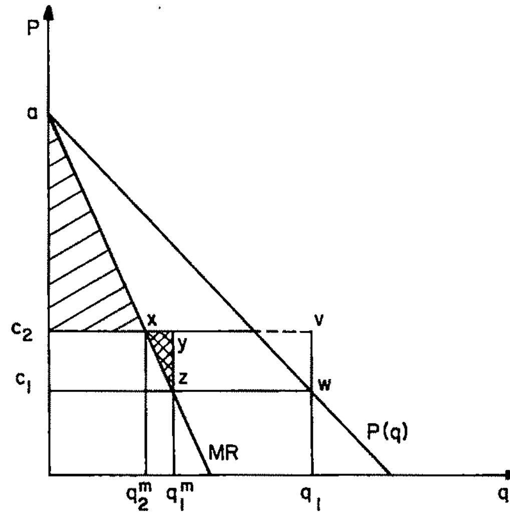


FIGURE 1. THE EFFECT OF A COST INCREASE FROM c_1 TO c_2.

Since $q_1^c > q_1^m$ (the duopoly output is at least as great as the monopoly output), a sufficient condition for this is $q_1^m > 2(q_1^m - q_2^m)$ or $q_2^m > q_1^m/2$. This condition holds for small changes in costs. In particular, it is satisfied unless the change in costs is so large that the optimal monopoly output is halved. Since it seems unlikely that a monopolist would prefer to halve its output at the current price in preference to changing the price, the result is quite general.

In some sense the result of this section is not surprising since, as Akerlof-Yellen argue, the cost from not changing one's price is of second order in the change in costs only if the profit function is differentiable with respect to price. For Bertrand duopolies the profit function is not differentiable, and indeed (2) is of first order in the change in costs while (1) is of second order. However, if we let the duopoly produce differentiated products, the profit functions become differentiable and both losses are of second order. Yet we show in the following section that as the two goods become better and better substitutes, the analysis in this section becomes more relevant.

It is important to note that while Bertrand duopolists respond more to changes in costs,

they respond less to changes in demand. With constant marginal costs the duopoly never changes its price when demand changes. On the other hand, apart from the exceptional case where the elasticity of demand is unaffected, the monopolist loses by not changing its price in response to a change in demand.

The analysis presented in this section has two shortcomings. First, the duopolists lose money in equilibrium. If they do not change their prices, the new equilibrium has $P_2 = c_2$, but they must incur the fixed cost of changing their prices. If they do not change their prices, they sell at a price less than marginal cost.

Second, the analysis does not carry over to the case of a cost decrease. In that case the new Bertrand equilibrium has $P_2 = c_2$. However, if each firm changed its price to that level it would lose f. Thus one firm can do better by not changing its price, selling nothing, and earning zero profits. It is therefore not a Nash equilibrium for both firms to decrease their prices to c_2. However, it is also not a Nash equilibrium for neither firm to change its price since one firm could profitably deviate by undercutting the price $P_1 = c_1$ slightly. The only equilibrium involves mixed strategies.[4]

Both of these shortcomings are due to the zero-profit nature of Bertrand competition. We show in Section II that if one allows for some degree of product differentiation, these problems disappear. Although the incentive for a duopolist to change its price is somewhat dampened with differentiated products since demand is less responsive to price differences, we show that duopolists may nonetheless change their prices more frequently than monopolists.

II. A Model with Differentiated Products

In Section I we showed that cost changes and demand changes have differing effects on the incentives of duopolists and monopolists to change their prices. Both costs and demand are affected when overall prices move, and it is such movements that are probably the main reason for price changes in the studies mentioned in the introductory section. Therefore, in this section, our focus is on the effects of changes in the overall price level.

We consider an industry in which two goods are produced. The demand for goods 1 and 2 is given by

$$(5)\quad q_t^1 = a/2 - (b/2 + d)P_t^1/S_t + dP_t^2/S_t$$
$$q_t^2 = a/2 - (b/2 + d)P_t^2/S_t + dP_t^1/S_t,$$

where a, b, and d are positive constants, S is the general price level, and $t = 1$ or 2 denotes the period. As can be seen from equation (5), the two goods are symmetric and d is a measure of their substitutability. The goods can be produced at constant marginal cost $S_t c$. Note that increases in S do not just raise costs, but also increase the quantity demanded at any price. This occurs because any given price now represents a smaller amount of real purchasing power. Therefore, profits deflated by S_t from producing good 1, π_t^1 are given by

$$(6)\quad \pi_t^1 = \left[a/2 - (b/2 + d)P_t^1/S_t + dP_t^2/S_t\right] \times \left(P_t^1/S_t - c\right),$$

and similarly for good 2.

If the firms simultaneously choose prices and behave noncooperatively, firm 1 chooses P_t^1 to maximize (6). The first-order condition is

$$(7)\quad P_t^1 = dP_t^2/(b + 2d) + aS_t/(2b + 4d) + cS_t/2.$$

[4] The symmetric mixed strategies are straightforward to calculate. Suppose each firm charges a price in excess of P with probability $F(P)$. If a firm lowers its price to P, it has the lowest price with probability $F(P)$ (the probability that its rival has a higher price). Therefore, a firm that lowers its price to P earns $(P - c_2)q(P)F(P) - f$. In equilibrium the firm must be indifferent between lowering its price to P and leaving it unchanged. This implies that $F(P) = f/(P - c_2)q(P) = f/(P - c_2)(a - bP)$ in the linear case. The mixed strategy involves not changing the price with positive probability and has no other mass points. The lowest price charged is that in which the firm earns f if it is the only firm charging that price.

The Nash equilibrium prices in period 1 if the firms expect S_1 to be equal to S_2 are then

$$P_1^1 = P_1^2 = [a + c(b+2d)]S_1/2(b+d).$$

It is useful to rewrite (6) as

$$(8)\quad \pi_t^1 = -(b/2+d)\big[P_t^1 - dP_t^2/(b+2d) - aS_t/(2b+4d) - cS_t/2\big]^2/S_t^2 + (b/2+d)\big[dP_t^2/S_t(b+2d) + a/(2b+4d) - c/2\big]^2.$$

Equation (8) decomposes π_t^1 into a term incorporating the first-order condition (7), and a term that is independent of P_t^1. Now suppose that S changes unexpectedly from S_1 to S_2. We then ask how big this change in S has to be in order to induce the firms to change their prices in the presence of a fixed cost to changing prices, f.

We first calculate the increase in firm 1's profits from changing its price from P_1^1 to P_2^1, assuming *that firm 2 does not change its price* ($P_2^2 = P_1^2$). We will show shortly that this gives a lower bound on the increase in firm 1's profits from changing its price. Notice that the second line of (8) is the same regardless of the price that firm 1 charges. The change in firm 1's profit if it changes its price is therefore

$$(9)\quad \Delta\pi^1 = -(b/2+d)\big[P_2^1 - dP_2^2/(b+2d) - aS_2/(2b+4d) - cS_2/2\big]^2/S_2^2 + (b/2+d)\big[P_1^1 - dP_2^2/(b+2d) - aS_2/(2b+4d) - cS_2/2\big]^2/S_2^2.$$

But notice that P_2^1 will be set equal to the price that maximizes π_t^1 given that firm 2 is setting P_2^2. Using (7), the first term is equal to zero. Thus we have

$$(10)\quad \Delta\pi^1 = (b/2+d)\big[P_1^1 - dP_2^2/(b+2d) - aS_2/(2b+4d) - cS_2/2\big]^2/S_2^2.$$

Using (7) and rearranging this gives

$$(11)\quad \Delta\pi^1 = (b/2+d)(\Delta S)^2/S_2 \times [a/(b+2d)+c]^2/4,$$

where ΔS is $S_1 - S_2$.

It is immediate from (10) that (11) gives a lower bound to the change in firm 1's profits from changing its price. This can be seen by noting that increases (decreases) in firm 2's price tend to increase (10) when S increases (decreases). (To see this, notice that if $P_1^2 = P_2^2$, then $P_1^1 = dP_2^2/(b+2d) + aS_1/(2b+4d) + cS_1/2$. If S increases, the right-hand side of this expression exceeds P_1^1. This difference is even greater if P_2^2 exceeds P_1^2. Similarly, if S decreases, the right-hand side is less than P_1^1. The difference between the left- and right-hand sides is then even greater if P_2^2 is less than P_1^2.) Thus when (11) exceeds f, a duopolist will always change its price.

Compare this with the situation for a monopolist who sells both products. To bias the argument against our case, we suppose that the monopolist can change both of its prices if it incurs the cost f. Algebra analogous to that above yields the result that the increase in a monopolist's profits from changing its price is

$$(12)\quad \frac{(\Delta S)^2}{S_2}\frac{b(a/b+c)^2}{4}.$$

The difference between (12) and (11), the monopolist's and duopolist's incentives to change price, is proportional to

$$(13)\quad b(a/b+c)^2 - (b/2+d) \times (a/(b+2d)+c)^2.$$

The derivative of (13) with respect to d is

$$[a/(b+2d)]^2 - c^2,$$

which is negative for d bigger than $(a - bc)/2c$. As d increases this derivative converges to the constant $-c^2$ so that, for d

sufficiently big, (11) exceeds (12) and duopolists change their prices in response to a smaller change in S. If one considers the example in which a equals 10, c equals 5, and b equals 1, then if d exceeds 7, there exists a change in the price level such that duopolists will change their prices whereas the monopolist will not.

The above analysis assumes that if only one of the firms (say firm 2) changes its price, that both firms nonetheless sell nonnegative quantities. However, it is clear from (5) that if $P_t^2 > P_t^1$, and d is sufficiently large, then q_t^2 can become negative. Substituting (7) into the second equation in (5), one can see that this problem does not arise for ΔS sufficiently small. If the increase in S is sufficiently small, firm 2 will still wish to produce a positive output at its best response to firm 1's "old" price. This is equivalent to a requirement that f not be "too large." As a numerical example, for the case in which $a = 10$, $c = 5$, and $b = 1$, if f is less than 10 percent of a duopolist's period 1 profits, this problem of being driven to a "corner" does not arise as long as $d < 90$.

We now turn to an interpretation of these results. An increase in S has two effects: it raises demand and costs at the current price. The simplified model of Section I provides the intuition for why the duopolists have a greater incentive to change their prices in response to a cost change. A monopolist that changes the prices of both products together (so that $P_t^1 = P_t^2$) faces aggregate demand with (negative) slope b as in the previous section. Referring back to Figure 1, the gain to changing price is thus given by the area xyz. The individual duopolist, on the other hand, faces demand with slope $b/2 + d$. As a consequence, the marginal revenue curve is flatter if $d > b/2$ and in that case the area corresponding to xyz is larger for the duopolist.

On the other hand, duopolists are less affected by the change in demand. If d is zero, each duopolist faces the same incentives in its market as a monopolist would. If the monopolist can change both of its prices by paying f, it will change its prices in response to a smaller change in demand. Moreover, as d increases the duopolist becomes even less concerned, until with $d = \infty$ demand stops mattering. This is because the duopolists face more elastic perceived demand curves. They are therefore less able to exploit increases in demand and, consequently, have less incentive to increase price.

It remains to explain why the cost effect dominates the demand effect (locally) when d is large. The reason is that, although second order, a small change in one duopolist's price can have a very large effect on the outputs (and hence the profits) of both duopolists if d is sufficiently large. Thus, if an arbitrarily small increase in S induces firm 1 to raise its price an arbitrarily small amount, firm 1 loses a large proportion of its sales to firm 2. It is the presence of this externality that makes the effect of a cost change for a duopolist qualitatively different from a demand change for a monopolist.[5]

III. Conclusions

In an industry subject to fluctuations in the firms' costs, it is more costly for each member of a tight oligopoly to keep its price constant than it is for a monopolist. The reverse is true for fluctuations in demand. When both costs and demand are subject to

[5] These results are broadly consistent with the simulations of Akerlof-Yellen (1985) and Olivier Blanchard and Nobuhiro Kiyotaki (1985). They compute the lost profit from keeping prices unchanged as a fraction of profits in the former case and as a fraction of revenues in the latter. Both show that in response to a small increase in the money supply that these fractions are higher, the higher is the elasticity of the demand facing firms. This is consistent with our paper insofar as our results also depend on duopolists having flatter perceived demand curves than monopolists. Yet this apparent similarity masks some important differences. First, comparing only the elasticity of demand across firms does not take into account that monopolists are different from individual oligopolists both in that they are larger and are subject to fewer strategic interactions. Second, insofar as monopolists have higher profits (or revenues) than oligopolists, considering only such ratios tends to make monopolists automatically appear to veiw fixed prices as less onerous. Finally, their simulations do not place firms in contexts in which general inflation (or, as in the 1930's, deflation) affects costs together with demand.

inflationary or deflationary shocks, the cost effect dominates. As a result, in the two period models we present, circumstances that lead a monopolist to change its prices would always encourage duopolists to do so as well, while the reverse is not true. In this conclusion, we point out a few caveats and possible extensions of the analysis.

First, our analysis has been concerned exclusively with the monopoly–duopoly comparison. Yet, Dennis Carlton (1986) as well as Stigler suggest that price rigidity is monotonic in concentration so that duopolies change their prices less often than three firm oligopolies and so on. The analysis of this paper can probably be extended to cover these cases as well. What was crucial in our analysis is that perceived demand curves become flatter as there are more competitors. This makes price changes more attractive because some of the benefits derive at the expense of competitors. Insofar as oligopolists with many competitors can reasonably be thought to have perceived demand curves that are more elastic (because there are better substitutes produced by competitors, for instance), they will change their prices more frequently.

Second, our analysis of the actual frequency of price adjustment applies strictly only in our two-period setting. An extension to a more general dynamic setting thus seems desirable. In some sense this extension should be straightforward; as the incentives for changing prices are bigger for oligopolies, we should observe them changing their prices more often. Unfortunately, when considering dynamic games between duopolists one must allow the strategies of the firms to depend on the history of their relationship. This considerably complicates the analysis. In particular, since price changes may precipitate price wars, there may be equilibria in which duopolies are reluctant to change their prices.[6]

Finally, the incentives to change price that firms face in this paper in response to exogenous changes in costs and demand, are related to the incentives that firms have to endogenously change costs and demand through innovation. Innovation on the cost and demand side corresponds to process and product innovation, respectively. To the extent that a product innovation cannot be appropriated by an individual firm but rather leads to a general change in industry demand, our model suggests that monopolists have a greater incentive to pursue such product improvements. With respect to process innovations, however, two effects operate to make the incentives for innovation greater for duopolists. First, as Kenneth Arrow (1962) has pointed out, such an innovation may be worth more in a competitive industry simply because of its greater output. Second, as our model suggests, each duopolist's fear that its rival will gain an advantage by innovating alone, provides a great incentive for each duopolist to innovate.

[6]For some dynamic models that use a framework capable of addressing these difficult questions, see Robert Gertner (1985) and Sheshinski and Weiss (1985).

REFERENCES

Akerlof, George A. and Yellen, Janet L., "A Near-Rational Model of the Business Cycle, with Wage and Price Inertia," *Quarterly Journal of Economics,* Supplement 1985, *100*, 823–28.

Arrow, Kenneth, J., "Economic Welfare and the Allocation of Resources," in Richard R. Nelson, ed., *The Rate and Direction of Inventive Activity,* Princeton: Princeton University Press, 1962.

Barro, Robert, "A Theory of Monopolistic Price Adjustment," *Review of Economic Studies,* January 1972, *39*, 17–26.

Blanchard, Olivier J. and Kiyotaki, Nobuhiro, "Monopolistic Competition and the Effects of Aggregate Demand," *American Economic Review,* September 1987, *77*, 647–66.

Bronfenbrenner, Martin, "Applications of the Discontinuous Oligopoly Demand Curve," *Journal of Political Economy,* 1940, *48*, 420–27.

Carlton, Dennis W., "The Rigidity of Prices," *American Economic Review,* September 1986, *76*, 637–58.

Gertner, Robert, "Dynamic Duopoly with Price Inertia," MIT, mimeo., December

1985.

Hall, R. L. and Hitch, C. J., "Price Theory and Business Behavior," *Oxford Economic Papers,* May 1939, *2*, 12–45.

Mankiw, N. Gregory, "Small Menu Costs and Large Business Cycles: A Macroeconomic Model of Monopoly," *Quarterly Journal of Economics,* May 1985, *100,* 529–39.

Means, Gardiner, "Industrial Prices and Their Relative Inflexibility," U.S. Senate Document 13, 74th Congress, 1st Session, Washington, 1935.

Primeaux, Walter J. and Bomball, Mark R., "A Reexamination of the Kinky Oligopoly Demand Curve," *Journal of Political Economy,* July/August 1974, *82*, 851–62.

______ **and Smith, Mickey C.**, "Pricing Patterns and the Kinky Demand Curve," *Journal of Law and Economics,* April 1976, *19*, 189–99.

Rotemberg, Julio, "Aggregate Consequences of Fixed Costs of Price Adjustment," *American Economic Review*, June 1983, *73*, 433–36.

Sheshinski, Eytan and Weiss, Yoram, "Inflation and Costs of Price Adjustment," *Review of Economic Studies,* June 1977, *44*, 287–304.

______ **and** ______, "Inflation and Costs of Price Adjustment: The Duopoly Case," mimeo., October 1985.

Simon, Julian L., "A Further Test of the Kinky Oligopoly Demand Curve," *American Economic Review,* December 1969, *59*, 971–75.

Stigler, George J., "The Kinky Oligopoly Demand Curve and Rigid Prices," *Journal of Political Economy,* October 1947, *55*, 432–49.

Sweezy, Paul M., "Demand Under Conditions of Oligopoly," *Journal of Political Economy,* August 1939, *47*, 568–73.

U.S. Bureau of Labor Statistics, *Wholesale Prices and Price Indexes*, 1929–37, Washington: USGPO.

Equilibrium Incentives in Oligopoly

By CHAIM FERSHTMAN AND KENNETH L. JUDD*

We examine the incentives that owners of competing firms give their managers. We show that, in equilibrium, each manager will be paid in excess of his decision's marginal profit in a Cournot-quantity game, but paid less than the marginal profit in a price game. In the Cournot case, deviations from profit maximization are reduced by ex ante cost uncertainty and increased by correlation in the firms' costs.

Orthodox economic theory treats firms as economic agents with the sole objective of profit maximization. Some have criticized this view of the firm as being simplistic, arguing that real firms may consistently strive toward a different goal. For example, William Baumol (1958) suggested sales maximization as a possible objective function of firms. Later, when economists more seriously considered the fact that the modern corporation is characterized by a separation of ownership and management, their attention focused on managerial objectives (see Herbert Simon, 1964; Oliver Williamson, 1964; Michael Jensen and William Meckling, 1976; and the principal-agent literature, such as Stephen Ross, 1973).

It is generally argued that a proper analysis of the firm's objective function should be based on the analysis of the owner-manager relationship. A manager's objective depends on the structure of the incentives that his owner designs to motivate him. Owners often index managerial compensation to profits, sales, output, quality, and many other variables. Even if we accept the traditional view that owners want to maximize profits, the incentive scheme they design may imply managerial incentives different from profit maximization. For a monopolistic firm, the owner-manager relationship can be described as a standard principal-agent problem. Such analyses have yielded rich insights into the structure of agents' incentives.[1] For example, Bengt Holmstrom (1977) showed that compensation in optimal contracts would likely use information other than final profits.

However, when we discuss oligopolistic markets, the individual owner-manager relationships must be examined within the context of *rivalrous* owner-manager pairings. More generally, whenever the profit accruing to a principal-agent pair depends on decisions that other rational agents make, the potential interactions must be considered. In this paper we examine the incentive contracts that principals (owners) will choose for their agents (managers) in an oligopolistic context, focusing on how competing owners may strategically manipulate these incentive contracts and the resulting impact on the oligopoly outcome. This analysis will yield different insights as to why managerial compensation contracts may not depend solely on realized profits, and also examine

*Department of Economics, Tel Aviv University, Israel, and Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Evanston, IL 60201, respectively. The authors would like to thank Steve Matthews, John Panzar, and Mark Satterthwaite for their comments. Professor Judd also gratefully acknowledges the financial support of National Science Foundation grant nos. SES-8409786 and SES-8606581, the Sloan Foundation, the Hoover Institution, and the Kellogg Graduate School of Management.

[1] The principal-agent approach (Ross, 1973; James Mirrlees, 1976; Bengt Holmstrom, 1977; Milton Harris and Artur Raviv, 1979; Roger Myerson, 1982; and many others) assumes that a principal chooses an incentive structure for agents which maximizes his welfare subject to information constraints and adequate compensation for the agents.

interactions between the structure of internal incentives within a firm and market structure elements external to the firm.

Even though it comes as no surprise that the strategic use of incentives can be important, little work has been done on the problem. Once one begins to think of incentives as strategic tools, it is clear that there may be value to the owner of distorting his manager's incentives away from maximizing the owner's welfare if the reaction of the owner's competitors is beneficial. In the case of a monopoly firm, the optimal incentive structure is obviously the regular principal-agent problem since there are no opponents and, in the absence of risk-sharing and asymmetric information considerations, such an owner will motivate his manager to maximize profits. In the case of an oligopolistic market, the optimal incentive structure is not so clear a priori. For example, Chaim Fershtman (1985) showed that nonprofit-maximizing firms may enjoy more profits than profit-maximizing firms in a duopoly. The strategic trade analysis of James Brander and Barbara Spencer (1983, 1985) is also an example of a "principal," in their case a government, distorting the incentives faced by an "agent," the local firms, in order to change the behavior of a competing "principal-agent" pair, a foreign government and its firms, in a fashion that advances the principal's objectives. Our analysis also expands on their insights on international trade policy by allowing uncertainty in critical parameters. More recent work by Brander and Tracy Lewis (1986) and Vojislav Maksimovic (1986) showed that a firm's owners may alter its capital structure in order to alter their incentives and the competitors' behavior.

In this paper we examine equilibrium incentive contracts in an oligopoly. We show that profit-maximizing owners will almost never tell their managers to maximize profits when each firm's managers are aware of other managers' incentives since each manager will react to the incentives given to competing managers. For example, if one firm's manager is told to maximize sales revenue instead of profit, he will become a very aggressive seller. Since his payoff is thereby affected, there will be a different equilibrium outcome in the competition among the managers. Also, the other managers' equilibrium behavior will be affected if they are aware of the firm's new incentive for sales maximization or learn of it through repeated play. This reaction in the competing firms' managers' behavior gives each owner an opportunity to be a Stackelberg leader vis-à-vis the other firm's managers when he determines his managers' incentives. We find that this interaction causes owners to twist their managers away from profit maximization even though the owners care only about profits.

We find, however, that the nature of the desired distortion critically depends on the nature of oligopolistic competition. In the case of Cournot-quantity competition, we prove that each owner wants to motivate his manager toward high production in order to get competing managers, who are aware of these incentives, to reduce their output. Therefore, in equilibrium owners will give a positive incentive for sales. On the other hand, if firms are selling differentiated products and compete in price, each owner will want his manager to set a high price, thereby encouraging competing managers to also raise their prices. Therefore, with price competition owners will pay managers to keep sales low.

This paper also determines the impact of uncertainty and heterogeneity on the oligopoly outcome. In the Cournot-quantity game, we find that the equilibrium outcome with incentive contracts is more efficient than the simple Cournot outcome not only because of the increase in output but also because the low-cost firm's share of output is greater. Furthermore, if the firms' costs are uncorrelated, *ex ante* uncertainty at the time the incentive contracts are written will reduce the deviation from profit maximization.

Even though our models will be specific, it will be clear that the idea of strategic manipulation of agents' incentives is of general interest. For example, we could similarly analyze a sales manager's decisions when he establishes incentives for his salespeople, and show that he overcompensates his salespeople at the margin if that will cause competing salespeople to work less.

Also, many of our results continue to hold when incomplete information and a moral hazard manager determine the information available for contracting purposes (see Fershtman and Kenneth Judd, 1987).

Section I describes the general framework we examine. Section II examines the results for a Cournot industry with random demand, whereas Section III examines the case of random costs. Section IV studies the case of n Cournot firms. Section V examines price competition in a differentiated product market. Section VI summarizes this study's results.

I. The Basic Model

Our model assumes two firms, each with an owner and a manager. When we say "owner," we mean a decision maker whose objective is to maximize the expected profits of the firm. This could be the actual owner, a board of directors, or a chief executive officer. "Manager" refers to an agent that the owner hires to observe demand and cost conditions and make the real-time decisions concerning output and/or price. While we will refer to the profit-maximizing agent as the owner, he in turn could be an employee who has been given incentives to maximize firm profits.

We examine a two-stage game. In the first stage, the owners of each firm simultaneously determine the incentive structure for its manager, knowing the true probability distributions governing demand and costs. Each owner must offer his manager a contract under which the manager expects to receive his opportunity cost of participation; at this stage the manager shares his employer's uncertainty about demand and costs and the belief about the incentives under which the opposing manager will work. In the second stage, the competing managers play an oligopoly game, with each firm's manager knowing his incentive contract and those of competing managers. In the second stage, the realized nature of demand and costs facing all firms will be perfectly known and common knowledge among the managers. After all sales have been made, each owner observes the costs and sales, and hence profits, of his firm.

Before continuing, we should note that our analysis is equivalent to another view of the market for managers.[2] Some will argue that instead of shareholders hiring managers, it is managers who propose incentive structures to the capital market, which then chooses among the competing managerial proposals. Even if one views the managers as making the first move, the resulting game is equivalent to our game as long as any contract which the managers can propose can also be proposed to the managers in our game, and vice versa. The order of who proposes the incentive contract, firm or potential managers, is not important. The crucial assumption is that the firm gets all the rent from the relationship, an outcome that will occur in either situation as long as there are a large number of potential managers per firm: if the firm proposes an incentive scheme in a take-it-or-leave-it fashion, it need offer a manager only his opportunity cost; whereas if there are many managers with similar opportunity costs making proposals to the shareholders, competition among them will leave the winner only with his opportunity cost, and in both cases an optimal scheme from each firm's point of view will be proposed and accepted. In some respects, this alternative formulation is attractive since a crucial assumption of our analysis is that the "owners" observe only profits and sales, and do not bother learning about the day-to-day details of the firm's operation, an assumption which is a plausible description of shareholders. In any case, we will stick with the more common theoretical structure of an owner proposing a contract and the manager responding.

The assumption that each firm's manager in stage two knows the other firm's manager's incentive contract and costs is a natural one in this context. We view the manager's contracts as being infrequently altered and in force for a substantial amount of time. Repeated play would presumably cause

[2] We thank an anonymous referee for pointing out this alternative view of the interaction between the capital and managerial markets.

managers eventually to learn one another's true incentives even if they were not initially common knowledge. However, despite this appeal to repeated play, we are assuming a single-shot game with common knowledge in stage two among managers about their incentives. A true repeated play specification of the managers' game would clearly generate many interesting new possibilities, but because of the intractable inference problems and the multiple-equilibria problems that arise in repeated games, it is beyond the scope of this paper to move beyond our two-period specification. Moreover, it will be clear in this two-stage game that each owner will want its manager's incentives to be common knowledge. For these reasons, we regard this critical information specification as appropriate and the two-period specification a reasonable one in which to study the issues on which we want to focus.

We assume that the incentive structure takes a particular form: risk-neutral managers are paid at the margin in proportion to a linear combination of profits and sales. More formally, firm i's managers will be given incentive to maximize

$$O_i = \alpha_i \pi_i + (1-\alpha_i) S_i,$$

where π_i and S_i are firm i's profits and sales.[3] This formulation is moderately general in that it is equivalent to maximizing linear combinations of profits and costs or sales and costs. We make no restrictions on α_i, allowing even negative values. We are assuming that, after the managers have acted and sales and production have been realized, the firms' owners can (or choose to) observe only profits and sales figures, not realizations of demand parameters or number of units sold. We allow managers to do whatever is in their best interest given their options and incentives, making the owner-manager relationship a delegation relationship, not a team relationship. The linearity restriction is not descriptively unreasonable. Furthermore, tractability demands that some restriction be put on the space of contracts since in similar generalized principal-agent problems it is known that equilibrium may not exist in unrestricted contracts (see Roger Myerson, 1982). While this is an unfortunate limitation of our analysis, it will be clear that it is not the reason for the qualitative nature of our results.[4]

Another crucial element of our model will be the assumption that there is uncertainty about crucial market parameters describing demand and costs at the time the incentives are determined. Such uncertainty from the owners' perspective is natural and also gives the managers a role as observers of these random variables. Uncertainty is also crucial to our focus on equilibria in which incentives are distorted away from profit maximization. We will argue that if we had no uncertainty about the *ex post* state of the market, then our analysis would be unconvincing since there would be no justification for ignoring quantity- or price-indexed contracts that would force the usual Cournot and Bertrand outcomes. However, simple deterministic forcing contracts will not be desired by owners when they face nontrivial uncertainty since each owner will want his manager to react to the eventual environment. Therefore, uncertainty is necessary to make the use of linear contracts in profits and sales reasonable and superior to contracts which yield the usual oligopoly outcomes.

The implicit restriction that i's manager's compensation depends on only firm i's sales and profits, not its competitor's, is motivated by a couple of realistic considerations. First,

[3] O_i will not be a manager's reward in general. Since his reward is linear in profits and sales, he is paid $A_i + B_i O_i$ for some constants A_i, B_i, with $B_i > 0$. Since he is risk-neutral, he acts to maximize O_i and the values of A_i and B_i are irrelevant.

[4] Recent work has indicated that the restriction to linear contracts is reasonable and does not mislead us. Bengt Holmstrom and Paul Milgrom (1987) show that linear contracts are optimal in some realistic continuous-time principal-agent problems. Fershtman and Judd (1987) showed that the basic insights of this paper continue to hold when moral hazard considerations also enter into the contracting problems of a duopolist. The focus on linear contracts here allows us to address questions that are intractable when we combine shirking by agents with a more complex dynamic structure.

a firm has much better information about its profits and sales than about its competitor's. Second, giving one's manager any incentive to increase a competitor's profits could possibly be illegal because of its clear role as a device to facilitate collusion. Third, even if we did allow cross effects in compensation, it will be clear that our main result of incentive manipulation would continue to hold true since each firm wants the other manager to operate in a cooperative fashion, but not its own manager.[5]

We examine the subgame-perfect Nash equilibrium of our two-stage game. In the second stage, we compute the Nash equilibrium that results when the firms' managers make simultaneous choices of their strategic variables, knowing one another's incentive contract and the realized nature of demand and costs. Below, we will examine cases in which the strategic variable is either price or quantity and make various assumptions concerning the information each firm has in the contract-writing stage about the eventual costs and demand. In the first stage, each owner simultaneously chooses its α_i, the relative weight it forces the manager to give to profits, with Nash equilibrium describing the outcome. In this game among the owners, each knows the payoff structure of each possible second-stage game as a function of the α's. We will refer to the stage-one equilibrium choice of the α_i and the resulting probability distribution of output and prices as the *incentive equilibrium*. We now move to the determination of incentive equilibria in several contexts.

II. Incentive Equilibrium with Cournot Competition and Random Demand

We first examine the issue of oligopolistic incentive structures for managers in a model of duopoly Cournot competition in a homogeneous good market. For reasons of tractability, demand is assumed to be linear:

$$p = a - bQ, \qquad a, b > 0, \tag{1}$$

where p is market price and Q is total output. q_i denotes the output of firm i, $i = 1,2$. Firm i will have constant unit cost $c_i \geqq 0$, $i = 1,2$. Both a or b are possibly unknown to all in stage one, but revealed to the managers at the beginning of stage two. We will make no special assumptions about the distribution of a and b other than assumptions on the support of their distributions necessary to assure that each firm's output will be positive in equilibrium. We see no reason to burden the reader with the extra algebra that would be needed when zero output is a possible equilibrium outcome, particularly since our interest is in the study of active oligopolies. The exact nature of such assumptions will be made explicit in the statements of the theorems below. In this section, c_1 and c_2 are known perfectly by all in both stages.

We solve for the incentive equilibrium in the standard backward fashion. In stage two, the manager of each firm observes a, b, c_1, c_2, α_1, and α_2, and chooses q_i to maximize O_i. In this case, O_i becomes

$$\begin{aligned} O_i = {} & \alpha_i(a - bQ - c_i)q_i \\ & + (1-\alpha_1)(a - bQ)q_i. \end{aligned} \tag{2}$$

Given α_1 and α_2, the Cournot reaction functions in quantity are

$$q_1 = \frac{a - bq_2}{2b} - \frac{\alpha_1 c_1}{2b}, \tag{3}$$

and symmetrically for firm two. Note that α_1 just affects the manager's perspective on costs. If $\alpha_1 < 1$, that is, firm one's manager moves away from strict profit maximization toward including consideration of sales, then firm one's reaction function moves out in a parallel fashion since the managers view $\alpha_1 c_1$ as the marginal cost of production. Therefore, as the owner of firm one changes α_1, he essentially changes his manager's reaction function. Symmetric results hold for firm two. These facts play the crucial role in the results below.

For values of α_i, $i = 1,2$, inherited from the outcome of stage one, stage-two equilibrium in terms of demand, cost, and

[5]Fershtman and Judd (1987) demonstrate this in a simple model.

incentive parameters is

$$p = (a + \alpha_1 c_1 + \alpha_2 c_2)/3, \tag{4a}$$

$$q_1 = (a - 2\alpha_1 c_1 + \alpha_2 c_2)/3b, \tag{4b}$$

$$\pi_1 = (a + \alpha_1 c_1 + \alpha_2 c_2 - 3c_1) \times (a - 2\alpha_1 c_1 + \alpha_2 c_2)/9b, \tag{4c}$$

and similarly for q_2. Note that as α_1 and α_2 are smaller, p is smaller, reflecting the fact that providing incentives for sales results in output beyond the profit-maximizing level.

Given the outcomes in stage two, firm one's owner chooses α_1 in stage one so as to maximize his expected profits net of his manager's opportunity costs. Since the cost of hiring a manager is fixed and unaffected by risk, this is equivalent to maximizing expected profits.[6] We first address the case in which only b is unknown *ex ante*. In this case, the reaction function for firm one's owner's choice of α_1 as a function of α_2 is

$$\alpha_1 = \frac{3}{2} - \frac{a}{4c_1} - \frac{\alpha_2}{4}\frac{c_2}{c_1}, \tag{5}$$

and similarly for firm two's owner's choice of α_2.

The case of uncertain b is particularly easy to examine since b is simply a parameter for the scale of the market and, as seen in (5), does not enter into the owners' choice of α_1 and α_2. Note that if firm two's manager maximizes profits, that is, $\alpha_2 = 1$, and costs are equal, then firm one's owner will choose $\alpha_1 < 1$. Therefore, profit-maximizing contracts generally do not arise in equilibrium. In fact, the final outcome is

$$\alpha_1 = 1 - \frac{a + 2c_2 - 3c_1}{5c_1}, \tag{6a}$$

$$q_1 = (2a - 6c_1 + 4c_2)/5b, \tag{6b}$$

and similarly for firm two. Since $\alpha_1 < 1$, if and only if $q_1 > 0$, we find that if both firms produce output, both will twist their manager's incentives away from strict profit maximization toward sales incentives as well.

The intuitive explanation for these results is given in Figure 1. For a particular b, R_i is firm i's reaction function, yielding q_i as a function of q_{3-i}. First, take the point of the owner of firm one. His choosing $\alpha_1 < 1$ pushes R_1 out and pushes the Nash equilibrium down firm two's manager's reaction function from A to B. The fact that α_1 is communicated to the manager of firm two means that firm one's owner acts as a Stackelberg leader with respect to firm two's manager. Here, however, each owner is a leader vis-à-vis his opponent's management. This dual leadership causes both owners to make their managers more aggressive sellers, leading both owners to choose an α_i less than unity, pushing both reaction curves out, and finally causing the stage-two Cournot equilibrium to shift from A to C. Theorem 1 summarizes.

THEOREM 1: *In a Cournot market, if a, c_1, and c_2 are known in stage one and both firms always produce positive quantities in equilibrium for any value in the support of b, then, for $i, j = 1,2$,*

$$\alpha_i = 1 - \frac{a + 2c_j - 3c_i}{5c_i}, \qquad i \neq j \tag{7a}$$

$$q_i = (2a - 6c_i + 4c_j)/5b, \qquad i \neq j \tag{7b}$$

$$p = (a + 2(c_1 + c_2))/5, \tag{7c}$$

implying that owners always give incentives for sales and may even penalize for profits if costs are sufficiently low.

There are several interesting comparisons between our incentive equilibrium outcome and the Cournot outcome. Total output in the incentive equilibrium always exceeds Cournot output, and profits and prices are lower. For example, if $c_1 = c_2 = c$, then Cournot price is $(a + 2c)/3$, whereas the

[6] Recall, from fn. 3, that i's manager is paid $A_i + B_i O_i$. Since managers are risk neutral, the owners can first set the optimal α's, yielding the O_i, and then promise A_i's, which set the means of $A_i + B_i O_i$ equal to the manager's opportunity cost. Hence, owners' profits equal expected profits minus managerial opportunity costs.

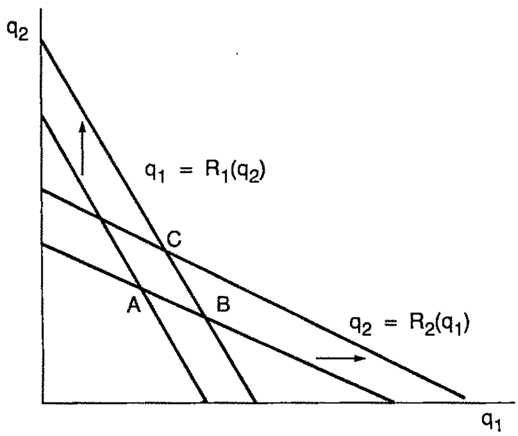


FIGURE 1

incentive equilibrium price is lower and equal to $(a+4c)/5$. Similarly, Cournot profits equal $(a-c)^2/9b$, whereas in our incentive equilibrium they are $2(a-c)^2/25b$, a lesser amount.

The incentive equilibrium outcome also has strong performance implications relative to the usual Cournot-quantity analysis. Since output is increased and oligopoly rents are lower, efficiency is improved. However, the incentive equilibria are more efficient not only because price is closer to marginal cost but also because production rises relatively more at the low-cost firm. If firm one is the low-cost firm, straightforward calculations show that its market share is $1+(c_2-c_1)/(a-2c_1+c_2)$ times greater in the incentive equilibrium than in the Cournot equilibrium. Corollary 1 summarizes.

COROLLARY 1: *Under the assumptions of Theorem* 1, *incentive equilibria in the quantity game generates greater output, lower rents, lower prices, and a more efficient allocation of production than the usual Cournot equilibria.*

The case of uncertain a is similarly examined. Let $\bar{a}$ denote the mean of a. Proceeding as above, we find Theorem 2.

THEOREM 2: *If b, c_1, and c_2 are known in stage one and if the minimum value of a with nonzero probability exceeds $2c_2-c$ and $2c_1-c_2$, then in the incentive equilibrium*

$$\alpha_1 = 1-(\bar{a}+2c_2-3c)/5c_1, \tag{8a}$$

$$q_1 = (2a-6c_1+4c_2)/5b, \tag{8b}$$

and similarly for firm two.

Before continuing, we should discuss one alternative formulation and its relationship to our game, particularly since it will give an argument as to why the addition of uncertainty was important to our analysis. Suppose that owners could write only contracts that force managers to produce a certain level of output and that there were no uncertainty in the level of demand. Such a world is equivalent to the usual Cournot game. Now suppose that the firms could write both these forcing contracts and the linear contracts we studied above. If firm one wrote a contract forcing its Cournot level of output, the best firm two could do would also be to write a contract that specifies its Cournot output since it could not manipulate the performance of firm one's manager. In Figure 1, such a forcing contract would cause R_1 to be vertical, a graphical representation of its nonmanipulability. Of course, the same argument applies to firm one. Therefore, the usual Cournot outcome is also an equilibrium if we assume no uncertainty and allow quantity-forcing contracts.

This observation does not, however, immediately eliminate the incentive equilibrium we computed above since it also remains an equilibrium in this extended game. The crucial fact is that, without uncertainty, a firm can choose any point along its opposing manager's reaction curve by choosing a quantity-forcing contract or a linear contract. If firm one believed that firm two was going to write the incentive equilibrium contract with its manager, then firm one is indifferent between writing a contract that forces its manager to produce the best point along firm two's manager's reaction curve and giving its manager the incentive equilibrium contract that will also produce that outcome. Similarly, if firm two's owner believed that one's manager was going to write the incentive equilibrium contract, it could do no better than to write its incentive equilibrium

contract. Therefore, multiple equilibria often result if both forcing and linear incentive contracts were possible.

In many cases of multiple equilibria, there is no reason to choose one over the other. However, the incentive equilibria would not be the natural one to focus on here in the absence of uncertainty. To argue this, we appeal to focal point considerations. Since our incentive equilibrium often results in less profit for both firms (and surely will if costs are identical), the incentive equilibrium would often be strictly Pareto inferior. In such cases, focal point considerations argue that the owners would realize that it is in their mutual interest to act according to the simple Cournot allocation implemented by forcing contracts. Therefore, the incentive equilibria lose much of their appeal in deterministic versions of our model.

However, if there are nontrivial levels of uncertainty, then such *noncontingent* quantity-forcing contracts would not be desirable since the owner would want the manager to be able to respond to contingencies that the owner does not observe, but which do affect his profits. Such flexibility could be partially attained in this context by a profit-maximizing contract. If both firms chose profit-maximizing contracts, then the state-contingent Cournot outcomes would result. However, once firms chose such contracts for their managers, each manager will react to deviations in the other's incentive contract. Therefore, by assuming uncertainty, we have both given a function to the manager and also increased the plausibility of our incentive equilibrium relative to one important perturbation of our game.

These comments also apply to the trade policy analyses in the papers by Brander and Spencer (1983, 1985). Their models will also have additional equilibria in similarly extended strategy spaces, with the extra equilibria sometimes being mutually preferable to both nations; however, uncertainty about the underlying profit opportunities will again make the linear contract equilibria the more plausible ones. The nature of our results also generalizes, implying that the strategic trade interventions will tend to be less valuable in the presence of uncertainty.

Section II has demonstrated the basic insight in our analysis: profit-maximizing owners may not want to give profit-maximizing incentives to their managers because an owner can influence the outcome of the competition between the managers in his favor by distorting his manager's incentives. This result does not rely on asymmetric information considerations as in Holmstrom, since a firm in this model will choose profit-maximizing contracts if it faces no competition. This result shows that internal relationships and incentives can be distorted and manipulated for interfirm strategic reasons, giving a new and fundamentally different role for internal contracts. In the following sections we elaborate on this theme for the cases of random costs, multiple-firm oligopoly, and price competition in differentiated markets.

III. Incentive Equilibrium with Cournot Competition and Random Costs

The case of random costs is substantially different. We examine it because new results concerning the impact of inter-firm heterogeneity are obtained.

Suppose that c_1 and c_2 are identically distributed with mean μ, variance σ^2, and correlation coefficient r. Let $v=\sigma/\mu$ be the coefficient of variation. Again, we will assume that the cost randomness is contained so that output for each firm is positive in each state of the world. We assume that each manager knows the other's costs in stage two. In this section, we assume that a and b, the demand parameters, are known perfectly in both stages. Therefore, the stage-two reaction functions are given by (3). In stage one, owner i chooses α_i to maximize expected profits given his expectation of α_{3-i}. Expected profits are given by *ex ante* expectation of (4c). Stage-one reaction functions are

$$(9)\quad \alpha_i=\frac{3}{2}-\frac{a}{4\mu}\frac{1}{v^2+1}-\frac{\alpha_{3-i}}{4}\frac{1+rv^2}{1+v^2},\qquad i=1,2.$$

Understanding the dependence of this reaction function on r, v, and μ is crucial to understanding the equilibrium results. If

there were no reaction by firm two's manager to firm one's incentive structure, there would be no gain to the owner from distorting his manager's incentives. The marginal gain to firm one's owner of increasing α_1 by $d\alpha_1$, assuming firm two's manager does *not* react to this change in his opponent's incentives, is

$$2(1-\alpha_1)(3b)^{-1}E\{c\}^2 d\alpha_1,$$

which is zero at $\alpha_1 = 1$, the profit-maximizing contract. However, since the manager of firm two will react in the stage-two equilibrium by increasing output as α_1 is increased, the marginal loss of increasing α_1 arising from this reaction is

$$\left(a\mu + \alpha_2 r\sigma^2 - 2\alpha_1(\mu^2+\sigma^2)\right)(3b)^{-1}d\alpha_1,$$

which is positive if q_1 is positive for all c_1, c_2 realizations. The reaction function chooses α_1, which equates the marginal gain and loss of an increase in α_1. As the variance, σ^2, increases, marginal losses due to deviations from profit-maximizing incentives increase, pushing the optimal α_1 toward 1. Also, if α_1 is near its optimal value given α_2, the gains from such deviations fall as σ^2 rises. Hence, we see that as σ^2 rises, firms move toward a profit maximization. Similarly, as costs are more correlated, the benefits of deviations from profit-maximizing incentives rise, implying that the optimal α_1 falls. Also, as a/μ rises, that is, the choke price rises relative to mean cost, the profit margin is greater and firms move away from profit-maximization incentives, as was the case for deterministic c.

Theorem 3 follows directly from an examination of the reaction functions.

THEOREM 3: *With ex ante uncertain and identically distributed costs, if q_1 and q_2 are nonnegative in equilibrium for all realizations of (c_1, c_2), then in equilibrium,*

$$\alpha_1 = \alpha_2 = \alpha = \frac{6(v^2+1) - a/\mu}{(4+r)v^2+5}. \tag{10}$$

Therefore, (i) α rises as a/μ falls and as v and r rise, and (ii) $\alpha < 1$ for r sufficiently close to 1 and v sufficiently close to zero.

The case of random costs is somewhat richer but more difficult to analyze completely. If the equilibrium α is less than unity, then an increase in the uncertainty of costs and their correlation will cause firms to move closer to profit maximization because it is more difficult to choose the right α conditional on the realized costs. We are not able to prove that α is always less than 1, but we know of no case in which it is not. The formula for α would seem to indicate that α could exceed one, but only if the variance of costs is large and costs are not perfectly correlated. This situation could possibly lead to negative output according to (6b) and violating the nonnegativity condition on output. To determine whether this occurs, one would have to impose specific distributions on the random variables. We want to confine the analysis in this study to cases in which examination of the random variables' first and second moments and weak conditions on their support is sufficient. Since the nonnegativity constraints on output are satisfied for r close to one or when the support for costs is small, yielding a v nearly zero, the formula for α in Theorem 3 is valid in these cases and (ii) of Theorem 3 holds, showing that Theorem 3 applies for a nontrivial set of cases.

This section shows how cost shocks affect the equilibrium nature of incentives. If cost shocks are commonly experienced, as in the case of an uncertain price for a common input, then the owners decide to distort incentives. However, if shocks are not commonly experienced, then deviations from profit maximization are reduced. Similarly, if there is too much variance in costs, then owners are not as willing to distort incentives away from profit maximization.

IV. Equilibrium with Many Firms

We saw above in a duopoly that owners may distort their managers' incentives if each firm's manager reacts to distortions in the competing managers' incentives. It is natural to ask next if these distortions of owners' incentives disappear as the industry is less

concentrated. We establish this formally in the case of perfectly correlated uniform costs. The same results for the cases of uncertain a and b are easily proven.

THEOREM 4: *As n approaches ∞, the firms' managers become profit maximizers, that is, $\alpha_i \to 1$, if a is uncertain, b is uncertain, or costs are equal but uncertain in stage one.*

PROOF:

Firm i's objective function is

$$(11) \qquad \alpha_i(a-bQ-c)q_i + (1-\alpha_i)q_i(a-bQ).$$

The first-order condition for choosing q_i is

$$(12) \qquad (a-2bq_i-b\overline{Q}_i)-\alpha_i c=0,$$

where $\overline{Q}_i = Q - q_i$. The second-order conditions are clearly satisfied. Thus, the ith firm's reaction function is

$$(13) \quad q_i = \frac{a-b\overline{Q}_i-\alpha_i c}{2b}, \quad Si=1,\dots,n.$$

Summing (13) yields

$$\sum_{j=1}^{n} q_j = \frac{1}{2b}\left(na-b(n-1)Q-\sum_{j=1}^{n}\alpha_j c\right).$$

Since $\Sigma q_i = Q$, then $b(n+1)Q = na - \Sigma\alpha_j c$. Therefore,

$$(14) \qquad Q=\frac{1}{b(n+1)}\left(na-c\sum\alpha_j\right).$$

Substituting (14) into (13) $(\overline{Q}_i = Q - q_i)$ yields

$$2bq_i = a-\alpha_i c-\frac{n}{n+1}a+\frac{c\Sigma\alpha_j}{n+1}+bq_i,$$

$$q_i=\frac{1}{b(n+1)}\left(a+c\sum_{j\neq 1}\alpha_j-nc\alpha_i\right).$$

From (14) we can calculate the price

$$p = a-\frac{1}{n+1}\left(na-c\sum_{j=1}^{n}\alpha_j\right) = \frac{1}{n+1}\left(a+c\sum_{j=1}^{n}\alpha_j\right).$$

The ith firm's expected profit when c is unknown in stage one is

$$(15) \quad \pi_i = \frac{1}{b(n+1)^2} \times E\left\{\left(a-c\sum_{j\neq i}\alpha_j-nc\alpha_i\right) \times\left(a+c\sum_{j=1}^{n}\alpha_j-(n+1)c\right)\right\}.$$

Maximizing the above profit functions for each i yields

$$(16) \quad E\left\{c\left(a+c\sum_{j\neq i}\alpha_j-nc\alpha_i\right) - nc\left(a+c\sum_{j=1}^{n}\alpha_j-(n+1)c\right)\right\}=0.$$

Since at the symmetric equilibrium $\alpha_i = \alpha$ for all i,

$$(17) \quad \alpha = 1-\frac{n-1}{n^2+1}\left(\frac{a\mu-\sigma^2-\mu^2}{1+\sigma^2+\mu^2}\right),$$

where $\mu = E\{c\}$ and σ^2 is the variance of c. Thus $\lim_{n\to\infty}\alpha = 1$. This proves Theorem 4.

This result is intuitively appealing because it coincides with our understanding of the perfectly competitive market. In the traditional theory of perfect competition with free entry, firms cannot afford to do anything other than be profit maximizers. If all firms have the same technology, the long-run equilibrium price is identical to minimum average cost. If one firm deviates from its

profit-maximizing output, its average cost is going to increase above the market price, implying that the firm loses money.

Since monopolists want their managers to maximize profits, we find that managers in both monopolized and competitive sectors will be told to maximize profits. Nonprofit-maximizing incentives will be given only in oligopolistic industries, showing that the relationship between market structure and managerial incentives will likely not be monotonic.

V. Price Competition and Incentive Equilibrium in a Differentiated Product Duopoly

The analysis of incentive equilibrium in a differentiated market is similar to the analysis in Section IV with one exception—now we assume price competition between firms selling differentiated products instead of Cournot-quantity competition. We assume that the demand is given by

$$q_i = A\tilde{\varepsilon} - bp_i + ap_{3-i}, \qquad i=1,2, \tag{18}$$

where $\tilde{\varepsilon}$ is a common shock to demand. We assume $\bar{\varepsilon}=1$. Also $b>a$, implying that the effect of a firm's own price on sales is greater than the effect of its rival's price. This is equivalent to concavity in the implicit linear-quadratic consumer utility function.

Owners know that the strategic variable in the competition between managers in the second stage is price. Thus, given an incentive structure which is a linear combination of profits and sales, firm i's manager will act so as to maximize

$$O_i = \alpha_i(p_i - c)(A\tilde{\varepsilon} - bp_i + ap_j) + (1-\alpha_i)p_i(A\tilde{\varepsilon} - bp_i + ap_j). \tag{19}$$

THEOREM 5: *When price is the strategic variable in the second stage of the competition among differentiated producers facing linear demand, $\alpha_i > 1$, that is, the incentive equilibrium is such that managers are overcompensated at the margin for profits.*

PROOF:

The reaction function of firm i's managers is given by

$$p_i = \frac{A + ap_j + \alpha_i cb}{2b}, \qquad i \neq j,\ i,j=1,2, \tag{20}$$

and the stage-two equilibrium prices, as a function of incentive and demand parameters, are

$$P_i(\alpha_i, \alpha_j, \tilde{\varepsilon}) = \frac{2bA\tilde{\varepsilon} + aA\tilde{\varepsilon} + a\alpha_j cb}{4b^2 - a^2} + \frac{2b^2\alpha_i c}{4b^2 - a^2}, \qquad j \neq i, \quad i,j=1,2. \tag{21}$$

Given the equilibrium in the second stage, the owners can compute their expected profits, π_i, $i=1,2$, as a function of the incentivestructures in their own firm as well as in the rival's firm:

$$\pi_i = E\left\{\frac{2bA\tilde{\varepsilon} + aA\tilde{\varepsilon} + a\alpha_j cb - 4b^2c + a^2c + 2b^2\alpha_i c}{4b^2 - a^2} \times \left[A\tilde{\varepsilon} - bP_i(\alpha_i, \alpha_j, \tilde{\varepsilon}) + aP_j(\alpha_i, \alpha_j, \tilde{\varepsilon})\right]\right\}. \tag{22}$$

By differentiating (22) with respect to α_i and equating it to zero, we find the reaction function of firm i's owner to firm $j=3-i$ to be

$$\alpha_i = m + \beta\alpha_j, \tag{23}$$

where

$$m = \frac{2ba^2A + a^3A - 6a^2b^2c + a^4c + 8b^4c}{4b^2(2b^2 - a^2)c},$$

$$\beta = \frac{a^3}{4b(2b^2 - a^2)}.$$

The equilibrium in the first stage of the game is a pair (α_1^*, α_2^*) such that (23) is satisfied for $i=1,2$. Substituting (22) into

(23) and solving for α_i yields

$$(24)\quad \alpha_i^* = 1 + \frac{(A-(a-b)c)(a^3+2a^2b)}{bc(8b^3-4a^2b-a^3)} > 1, \quad i=1,2$$

since $a < b$. The overcompensation for profits can also be interpreted as an owner's tax on the manager for his input expenditures. This tax disciplines the manager and prevents him from being too aggressive in his pricing strategy.

An immediate corollary of Theorem 5 is that the price in the incentive equilibrium is above the equilibrium price in an industry in which managers maximize profit, the usual specification of behavior in a differentiated market. This can be illustrated by a reaction function analysis. The crucial difference between this case and the Cournot-quantity case is that here the reaction curves in prices slope upward, that is, the greater a firm's expectation about its opponent's price, the greater will be the price he chooses. Under the profit-maximization hypothesis the equilibrium prices are (p_1^*, p_2^*) in Figure 2. By penalizing managers on sales at the margin, as occurs here since the equilibrium of the owner's game implies that $\alpha > 1$ for both firms, managers will price less aggressively than under the regular profit-maximization hypothesis. This pushes their reaction functions upward, a shift evident from (20) which describes i's reaction function. This mutual restraint results in equilibrium moving outward, a direction favorable to both, and leading to prices of $(\hat{p}_1, \hat{p}_2)$, which are higher than the equilibrium prices under the profit-maximization hypothesis.

Comparing the above result with the equilibrium in the quantity competition case[7]

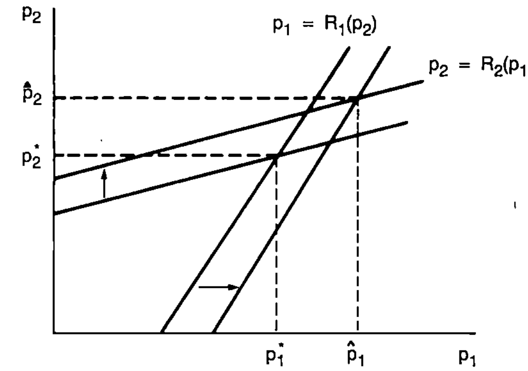


FIGURE 2

demonstrates that the equilibrium incentive structure depends on the way firms compete in the market. In the quantity competition case, $\alpha < 1$ and owners motivate managers to behave aggressively and to produce beyond the profit-maximization level, whereas in the price competition case $\alpha > 1$ and managers behave nonaggressively. In the quantity competition case each owner, acting as a Stackelberg leader with respect to the opposing manager, recognizes the negative slope of its rival manager's reaction function and therefore wants his manager to expand output. In the price competition case, each owner knows that any credible increase in its own price will be followed by an increase in its rival's price, therefore motivating its manager to be less aggressive and charge a price above the profit-maximizing price.

Moreover, the performance implications of incentive equilibrium differ in the two cases. In the quantity competition case, the incentive equilibrium increases efficiency and reduces oligopoly rents since the outcome is closer to perfect competition than the outcome in the regular Cournot competition. However, in the price competition case, incentive equilibrium essentially pushes the price toward the monopolistic price. The structure of our incentive contract game would therefore be one of mutual advantage to the firms and hence one toward which they would like to move instead of the decision-making structure implicit in the usual Bertrand analysis.

[7]Had we assumed our differentiated producers competed in quantities, then the results would have resembled those of the nondifferentiated Cournot analysis. Therefore the comparisons we make here result from the mode of competition, not from product differentiation. The product differentiation feature was added to the analysis of Section IV in order to keep price competition from always resulting in marginal cost pricing.

Finally, the argument that incentive equilibria in the deterministic quantity competition case with no uncertainty were not plausible for focal point reasons when we expand the space of possible contracts does *not* apply here. Since firm profits are increased by incentive contracts, such considerations argue in favor of the incentive contracts over forcing contracts (which here would be interpreted to force the manager to choose a certain price) even when both were equilibria in the expanded contract space. These arguments indicate that our theory of incentive equilibria may be more relevant for the case of differentiated markets.

VI. Concluding Comments

This paper has examined the interactions of internal contracting and external strategic considerations. We found a principal (firm owner) will want to distort the incentives of his agents (firm managers) in order to affect the outcome of the competition between his agent and competing agents. The general implications of our analysis are clear. In general, the owner of a firm will alter his managers' incentives in that direction which will cause opposing agents to change their behavior in beneficial directions. For example, if advertising will cause opposing firms to reduce their advertising, then a firm's owner will give his managers extra incentive to advertise. This can be implemented by explicit incentives or by hiring agents who are known to be inclined to aggressively advertise. In some cases, various asymmetries may cause the owners to distort their managers' behavior in opposing directions. For example, if in the differentiated products case firm one is a price setter, but firm two, for some technical reason, fixes his quantity, then firm one's manager will be paid to overproduce in order to get firm two's manager to reduce his output, but firm two's manager may be paid to keep his price high and output low in order to encourage firm one's manager to allow the market prices to be high. The variety of problems that can be analyzed by focusing on this joint determination of internal incentives and external environment is obvious.

There are a variety of directions which further research should pursue. The major weakness of the analysis above is the assumption of linear contracts and the absence of a detailed asymmetric information structure which motivates the existence of contracts in the first place. This study is offered as an imperfect but intuitive and suggestive analysis of the possibilities that arise when we jointly examine managerial incentives and market structure. A more recent paper by the authors (Fershtman and Judd, 1987) examines a model with a more standard incomplete information and moral hazard structure, which demonstrates that the intuitive results derived above continue to hold within a more standard principal-agent structure.

We also assumed that the managers play a simple Nash noncooperative equilibrium when they compete. An alternative theory of their behavior would be to have them bargain toward some outcome that is cooperative from their point of view. While this may substantially affect the outcome of the managers' game for any given set of managerial incentives, the owners would still take into account the impact their decisions have on the outcome of the managers' decision-making process. For example, if the managers were known by the owners to bargain in accordance with a "split the gains from trade" rule, then each owner will strive to increase his profits by demanding a large bond from the manager and then contract to give a large portion of it back whether bargaining succeeds. This will raise the manager's threat point, making the agreement point more favorable to his firm, and increase the owner's profits. In any case, strategic manipulation of managerial incentives will be valuable to owners as long as the manager's incentives affect the joint allocation of profits in the manager's game, a feature which appears in most cooperative modes of interaction as well as noncooperative.

This paper has demonstrated that competing firms' owners will often distort their managers' objectives away from strict profit maximization for *strategic* reasons. This initial analysis made several simplifying as-

sumptions including linear payoffs to managers, absence of the usual moral hazard problems, and linear demand. Further research should generalize our analysis. However, it is clear from the basic intuition that distortions of managers' incentives are potentially important strategic instruments for owners of competing firms and point to the importance of external competitive conditions for the determination of internal relationships.

REFERENCES

Baumol, William, "On the Theory of Oligopoly," *Economica*, August 1958, *25*, 187–98.

Brander, James A. and Lewis, Tracy R., "Oligopoly and Financial Structure: The Limited Liability Effect," *American Economic Review*, December 1986, *76*, 956–70.

______ **and Spencer, Barbara J.,** "Export Subsidies and International Market Share Rivalry," *Journal of International Economics*, February 1985, *18*, 83–100.

Fershtman, Chaim, "Internal Organizations and Managerial Incentives as Strategic Variables in Competitive Environment," *International Journal of Industrial Organization*, *3*, 1985, 245–53.

______ **and Judd, Kenneth L.,** "Strategic Incentive Manipulation in Rivalrous Agency," Hoover Institution Working Papers in Economics E-87-11, 1987.

Holmstrom, Bengt, "On Incentives and Control in Organizations," unpublished doctoral dissertation, Stanford University, 1977.

______ **and Milgrom, Paul,** "Aggregation and Linearity in the Provision of Intertemporal Incentives," *Econometrica*, March 1987, *55*, 303–28.

Harris, Milton and Raviv, Artur, "Optimal Incentive Contracts with Imperfect Information," *Journal of Economic Theory*, April 1979, *20*, 231–59.

Jensen, Michael C. and Meckling, William H., "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics*, October 1976, *3*, 305–60.

Maksimovic, Vojislav, "Optimal Capital Structure in Oligopolies," unpublished doctoral dissertation, Harvard University, 1986.

McGuire, J., Chiu, J. and Elbing, A., "Executive Income, Sales, and Profits," *American Economic Review*, September 1962, *52*, 753–61.

Mirrlees, James, "The Optimal Structure of Incentives and Authority Within an Organization," *Bell Journal of Economics*, Spring 1976, *7*, 165–81.

Myerson, Roger, "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," *Journal of Mathematical Economics*, June 1982, *10*, 67–81.

Ross, Stephen, "The Economic Theory of Agency: The Principal's Problem," *American Economic Review*, May 1973, *63*, 134–39.

Simon, Herbert, "On the Concept of Organizational Goal," *Administrative Science Quarterly*, June 1964, *9*, 1–21.

Spencer, Barbara J. and Brander, James A., "International *R&D* Rivalry and Industrial Strategy," *Review of Economic Studies*, October 1983, *50*, 707–22.

Williamson, Oliver E., *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*, Englewood Cliffs: Prentice-Hall, 1964.

The Simple Analytics of Competitive Equilibrium with Multiproduct Firms

By GLENN M. MACDONALD AND ALAN SLIVINSKI*

A model of free-entry perfectly competitive markets in which firms may produce more than one product is developed. The formulation is very simple and closely parallels the classic single-product model, the goal being to provide a useful, accessible tool for applications-oriented research. Examples of the kind of analysis made possible by the model, and simple extensions of it, are presented.

The familiar model of free-entry, competitive equilibrium has long played a central role in applied analysis of product markets. Among its stylizations is the restriction that firms produce a single output. For many situations, however, it is necessary to relax this assumption. While models permitting multiproduct firms exist, so far there is no framework that begins to rival the single-product analysis in terms of the ease with which it may be manipulated and extended to deal with specific applications. This paper develops such a model.[1]

Specifically, in Section I a two-good model is set out that parallels the classic single-product analysis very closely. The sole difference is that alongside the single-product ("specialized") technologies, which would usually be permitted, a multiproduct ("diversified") technology is available. The model's competitive equilibrium is then characterized, and it is shown that this equilibrium may take one of three forms. Obviously, if diversification offers large cost advantages, no specialized firms can operate in equilibrium, and conversely if there are sizable disadvantages. The only other possibility is that diversified firms and exactly one type of specialized firm operate contiguously.

Section II shows that the model is easy to manipulate, extends straightforwardly, and simple as it is, offers some new propositions. This demonstration involves examining the predictions the model offers in both its most general form and several extensions. To illustrate, basic features of equilibrium in the standard single-product environment are that price is determined by the cost of production with all operating firms producing the same level of output. Also, demand variation has no effect on either price or the actions of these firms. In contrast, in any equilibrium

*Centre for Decision Sciences and Econometrics, Social Science Centre, University of Western Ontario, London, Ontario, N6A 5C2. Comments from David Donaldson, Ignatius Horstmann, Boyan Jovanovic, Peter Lloyd, Michael Parkin, Charles Plott, Edward Prescott, Sherwin Rosen, Christopher Robinson, Hugo Sonnenschein, and the referees are gratefully acknowledged.

[1]Research allowing multiproduct firms comes in a variety of forms. Early work by Roy G. D. Allen, 1938; John Hicks, 1939; Paul Samuelson, 1947; as well as more recent efforts by Keith Laitinen, 1980, analyzed in detail the isolated behavior of firms having access to an m-input/n-output production technology. Also, the Arrow-Debreu-McKenzie general-equilibrium model allows each producer a distinct production set, so that firms might choose to produce many goods. More recently, the "contestability" literature (surveyed by Elizabeth Bailey and Ann Friedlander, 1982) analyzes a setting in which firms produce more than one good. Of all the contestability material, the work of William Baumol et al., 1982, Ch. 9, is the most closely related to the present analysis. Therein firms are permitted to choose a set of goods to produce, and a condition is provided that is necessary and sufficient for the (otherwise exogenously imposed) symmetric outcome to be supported in equilibrium. That this condition is indeed a relevant restriction is shown by means of a two-good numerical example in which the condition fails and the equilibrium is asymmetric. Finally, there is what might be termed the "where there is sawdust there may be 'pressed logs'" approach, dating back at least to Alfred Marshall, 1920, pp. 321–22, in which joint products are the result of unstructured technological complementarities.

involving multiproduct firms, identical actions on the part of all firms within an industry, and independence of these actions from demand, are inconsistent. That is, if all firms produce in the same fashion, demand plays a role in determining what firms do. Or, if demand does not play such a role, then firms cannot all make the same choices.

Comparative statics-type results are also explored. The model's main ingredients are demand functions, and fixed and variable costs. The effect of varying these entities is examined for either the general model or an extension, depending on where the clearest and most useful results lie. It is shown, for example, that when both diversified and specialized firms operate, the specialized firms play an "industry fringe" role, absorbing all variations in demand for the good they produce. One extension models the structure of fixed costs in terms of different kinds of substitutable public inputs, such as upper-level management (whose decisions affect the entire firm) and lower-level supervisors (where decisions affect only the production process for one good), and shows that when both diversified and specialized firms operate, the former employ more upper-level management overall and fewer supervisors on the process yielding the good produced by both types of firms.

Throughout, while the analysis is rigorous, an attempt is made to present the material in a fashion that accommodates the basic goal of producing a tool for applications-oriented research. As such, most of the formal analysis is suppressed.[2]

I. Analysis of the General Model

A. *A Single Product*

It is useful to begin by reviewing the manner in which a free-entry competitive equilibrium in the market for a single good is characterized. First, consider the supply side. Firms maximize profits $\pi \equiv pq - F - c(q)$, where p is the price of the good, q is the level of output, F is a fixed cost, and $c(q)$ is a variable-cost function, assumed to be increasing and strictly convex with $c''(q)$ continuous. Market demand is $Q = D(p)$, assumed to be continuous and monotonically decreasing in p. Letting N be the number of operating firms, free-entry competitive equilibrium values of N, q, and p must satisfy[3]

$$(1) \qquad p - c'(q)\begin{cases} = 0 & \text{if } q > 0 \\ \leq 0 & \text{if } q = 0, \end{cases}$$

$$(2) \qquad pq - F - c(q)\begin{cases} = 0 & \text{if } N > 0 \\ \leq 0 & \text{if } N = 0, \end{cases}$$

$$(3) \quad \text{and} \qquad D(p) = Nq.$$

(1) states that the choice of output must be profit maximizing—equating marginal cost to price if positive output is chosen, and marginal cost at zero output equaling or exceeding price otherwise. (2) requires profits to be zero if firms operate and not positive otherwise. (3) states the equality of market demand and total output.

Inclusion of the inequalities in (1) and (2) may strike the reader as pedantic. However, in the multiproduct case studied below, analogous (especially to (2)) inequalities play an important role, so that placing some emphasis on them at this point proves helpful. In the single-product setting, assuming $c'(0) = 0$ guarantees (1) may always be treated as an equality. Supposing $N > 0$, (2) is also an equality. (1) and (2) together then give the familiar result that if $N > 0$, the equilibrium product price p^* and firm output q^* are determined solely by the supply side: q^* is the rate of output that equates marginal and average cost, and p^* is the implied value of

[2] More detail is available in the authors' 1986 paper.

[3] The number of firms is treated as a continuous variable. Doing so can be justified in a number of ways, but, in any event, nothing of consequence in the analysis depends on this approach. The same route will be taken in the multiproduct analysis.

average cost.[4] N^* is then determined by the requirement that $N^*q^* = D(p^*)$.

Finally, for later reference, it is worth noting the simple result that if two technologies are available for producing any single commodity, in equilibrium only the one having the lower minimum value of average cost will be used. The equilibrium is then precisely as above, unless the minimum levels of average cost for the two technologies coincide, in which case firms are indifferent as to which they employ, and the equilibrium "pattern of technology" is indeterminate.

B. *Multiproduct Model*

The approach taken in the multiproduct setting follows the single-product logic as closely as possible. There are two goods, called α and β, and indexed by j; $j = \alpha, \beta$. To avoid repetition, k will be whichever of α, β is "not j." Demand for good j will be written $Q_j = D_j(p_\alpha, p_\beta)$, where p_j is the price of good j. $D_j(p_\alpha, p_\beta)$ is assumed continuous, decreasing in p_j and increasing in p_k. Because cases where the market does not operate are not of interest, it will be assumed that $D_j(\cdot)$ is positive for both goods irrespective of prices.

Any firm may produce either both goods or just one. If it produces only good j, it will be called a "type j" firm; otherwise it will be referred to as "diversified." A type j firm faces total costs

$$F_j + c_j(q_j),$$

where F_j is a fixed cost and $c_j(q_j)$ an increasing and strictly convex variable-cost function with c_j'' continuous.

[4]Strictly, it is necessary to impose a condition that guarantees there is some finite q for which $c'(q) = [F + c(q)] \div q$. To obtain the condition, note that for some $\hat{q} \in (0, q)$, $F + c(0) = F + c(q) - c'(q)q + c''(\hat{q})q^2/2$. Thus $[F + c(q)]/q - c'(q) = [F + c(0)]/q - c''(\hat{q})q/2$. The right-hand side is continuous in q for $q \in (0, \infty)$, positive for q sufficiently small, and negative for q sufficiently large provided $c''(\hat{q}) \cdot q$ is bounded away from zero. Thus, a sufficient condition for the existence of the q in question is $c''(q) > \eta$ for all q, where $\eta > 0$ is a constant.

Total costs for a diversified firm are

$$F + c(\tilde{q}_\alpha, \tilde{q}_\beta),$$

where $\tilde{q}_j$ is the output of product j (the tilde indicating that the output is produced by a diversified firm), F is a fixed cost, and $c(\tilde{q}_\alpha, \tilde{q}_\beta)$ is the variable-cost function. $c(\cdot)$ is assumed to be increasing and strictly convex with continuous second partial derivatives.

Given product demands and the structure of costs, competitive equilibrium in the multiproduct setting involves satisfaction of a collection of relationships analogous to (1)–(3). Endogenous variables to be determined are: (i) product prices, p_j; (ii) output levels for type j firms, q_j; (iii) output levels for diversified firms, $\tilde{q}_j$; (iv) the number of type j firms, N_j; and (v) the number of diversified firms, N. The required conditions are: For $j = \alpha, \beta$,

$$(1') \quad \begin{cases} p_j - c_j'(q_j) & \begin{cases} = 0 & \text{if } q_j > 0 \\ \leq 0 & \text{if } q_j = 0 \end{cases} \\ p_j - \dfrac{\partial}{\partial \tilde{q}_j} c(\tilde{q}_\alpha, \tilde{q}_\beta) & \begin{cases} = 0 & \text{if } \tilde{q}_j > 0 \\ \leq 0 & \text{if } \tilde{q}_j = 0, \end{cases} \end{cases}$$

$$(2') \quad \begin{cases} p_j q_j - F_j - c_j(q_j) & \begin{cases} = 0 & \text{if } N_j > 0 \\ \leq 0 & \text{if } N_j = 0 \end{cases} \\ \displaystyle\sum_{j=\alpha,\beta} p_j \tilde{q}_j - F - c(\tilde{q}_\alpha, \tilde{q}_\beta) & \begin{cases} = 0 & \text{if } N > 0 \\ \leq 0 & \text{if } N = 0, \end{cases} \end{cases}$$

and

$$(3') \qquad D_j(p_\alpha, p_\beta) = N_j q_j + N\tilde{q}_j.$$

(1′)–(3′) may be interpreted in exactly the manner in which (1)–(3) were.

C. *Competitive Equilibrium in the Multiproduct Model*

The analysis of (1′)–(3′) can proceed without further assumptions. However, as in the single-good environment, consideration of uninteresting cases can be eliminated by imposing some simple conditions on the cost functions. In the single-good model, the con-

dition $c'(0)=0$ allowed (1) to be treated as an equality, in which case the equilibrium-product price and level of output per firm were determined independently of demand, assuming that some firms operate. In the multiproduct setting, two analogous conditions are employed: for $j=\alpha,\beta$,

$$(4) \quad c_j'(0)=0,$$

and

$$(5) \quad \tilde{q}_j=0 \Rightarrow \frac{\partial}{\partial \tilde{q}_j} c(\tilde{q}_\alpha, \tilde{q}_\beta)=0 \; \forall \; \tilde{q}_k$$

The new condition, (5), states that for all $\tilde{q}_k$, the marginal cost of producing good j is low for small $\tilde{q}_j$. Imposition of (4) and (5) permits (1′) to be treated as a system of three equalities, irrespective of which types of firms operate.[5]

If the analogy with the single-product setting were complete, it would now be assumed that $N>0$ and both $N_j>0$. Then (1′) and (2′) would be a set of equalities determining product prices and rates of output independently of demand. However, treated as equalities, (1′) and (2′) include four profit-maximization conditions and three zero-profit conditions. Yet, there are only six endogenous entities to be determined—q_j, $\tilde{q}_j$, and p_j, for $j=\alpha,\beta$. It follows that except for a "knife-edge" situation (described below), at most six of the relationships in (1′) and (2′) can hold as equalities. Because, under (4) and (5), (1′) always consists of equalities, the inequalities must be part of (2′), meaning at most two of the nonpositive profit conditions can be fulfilled simultaneously as equalities. The first conclusion is therefore that at most two of N^*, N_α^*, and N_β^* can be positive in equilibrium, where asterisks denote equilibrium values. If just one is positive, it must be N^*, for otherwise demand for one good would necessarily be unsatisfied. Consequently, there are three "configurations" to be analyzed: I. *Pure Specialization*—$N_\alpha^*>0$, $N_\beta^*>0$, and $N^*=0$; II. *Pure Diversification*—$N^*>0$ and $N_\alpha^*=N_\beta^*=0$; and III. *Mixed Production*—$N^*>0$ $N_j^*>0$, and $N_k^*=0$, $j=\alpha,\beta$. In what follows, these equilibria will be referred to as type S, D, or M_j (when $N_j>0$), respectively.

The intuition underlying this "configuration" result has much in common with the logic behind the proposition, in the one-good case, that only one technology can be used in equilibrium, apart from a suitable coincidence. Suppose all three types of firms did produce in equilibrium. Profit maximization and zero profit, for both types of specialized firms in particular, imply that p_j^* equals minimum average cost for good j in those firms. The conditions in (1′) and (2′) that refer to type j firms operate precisely as in the one-good case. Turning to diversified firms, their being part of this equilibrium requires that when they choose output levels to maximize profit given the prices just described, the multiproduct technology is such that they happen to earn exactly zero profits: a very special coincidence analogous to two different technologies yielding the same minimum average cost in the one-good model.

Pure Specialization is the case analyzed in every undergraduate microeconomics text, while Pure Diversification has been treated, at least partially, in various multiproduct-firm models. The Mixed Production configuration does not appear to have been explored in detail before. It is most interesting of all in that it permits firms to undertake different activities in an environment where all have identical access to technology—a feature shared only by Pure Specialization—but does not allow markets to be analyzed separately—a characteristic in common only with Pure Diversification. Moreover, the heterogeneity that exists in the Mixed Production setting enriches the descriptive accuracy of the analysis. Finally, the most useful and informative elaborations considered in Section II involve analysis of the Mixed Production configuration.

A second result, that is itself unremarkable but useful in what follows, parallels the single-good case more closely. Suppose one of the three possible configurations given

[5]As in the single-product setting, it is necessary to impose a condition guaranteeing that the multiple-output analogue of everywhere declining average cost does not occur. (6), imposed below, is sufficient.

above is the equilibrium outcome for some demands $D_j(p_\alpha, p_\beta)$. Then consider increasing the quantity of each good demanded at the equilibrium prices by the same factor, λ. From (1′)–(3′) it is immediate that the only adjustment implied by this change is that the number of firms of each type is raised by the factor λ, precisely as occurs in the single-good case. Therefore, all aspects of the equilibrium apart from N^* and N_j^* depend on the structure of demand only via "relative demand" $D_j(p_\alpha, p_\beta)/D_k(p_\alpha, p_\beta)$.[6] For future reference, a special "base" case is that in which relative demand depends only on relative prices, as would occur if, for example, all consumers had identical homothetic preferences.

D. *A Useful Diagram*

Given the three possible types of configurations, two questions are immediate. First, under what conditions do the different possibilities emerge? Second, what are the basic features of each? Most of the results may be represented in a single diagram, the development of which follows.

In constructing the model's equilibrium, a good deal of tedious discussion is avoided if two simple conditions are added. The first assumption is that for $j = \alpha, \beta$

$$(6) \qquad \tilde{q}_j \frac{\partial^2 c}{\partial \tilde{q}_j^2} + \tilde{q}_k \frac{\partial^2 c}{\partial \tilde{q}_j \partial \tilde{q}_k} > \varepsilon$$

for some constant $\varepsilon > 0$. (6) implies that if $\tilde{q}_j$ and $\tilde{q}_k$ are increased in the same proportion, the marginal cost of producing either rises, and does so nonnegligibly. The important part is that the marginal costs rise; the "nonnegligibly" portion is a technical issue.[7] An important "base case" is that for which such proportional increases leave relative marginal cost unchanged.

[6] Note that it is implicitly assumed that demands are large enough that the number of firms producing good j, $N^* + N_j^*$, is sufficient to warrant the assumption of competition. This assumption is exactly as in the classic single-product analysis. Of course, neither model offers any help on the question of how many firms are enough.

[7] See fn. 5.

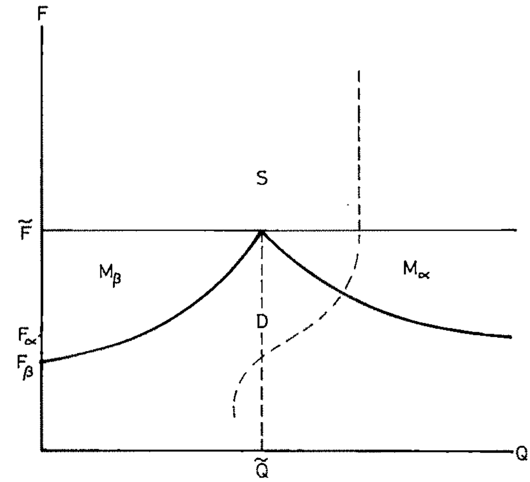


FIGURE 1

The second condition is: for $j = \alpha, \beta$

$$(7) \qquad \tilde{q}_k = 0 \Rightarrow c(\tilde{q}_\alpha, \tilde{q}_\beta) = c_j(\tilde{q}_j).$$

The difference between a diversified firm and a type j firm is that by incurring the fixed cost F, as opposed to F_j, the diversified firm has the "capacity" to produce $\tilde{q}_k$ as well as $\tilde{q}_j$. (7) states that if this capacity goes unexercised, diversified and type j firms have the same variable costs.

Under (6) and (7), the salient features of the model's equilibrium may be depicted as in Figure 1.

The diagram is drawn in the (Q, F) plane, where $Q \equiv Q_\alpha/Q_\beta$. Q is the relevant variable because, as established above, changes in demand that hold Q fixed for given prices affect only the absolute numbers of firms of each type, leaving prices, quantities per firm, and the configuration of firms unchanged.

Ignoring the dashed curve for the moment, the diagram partitions the (Q, F) plane into four regions. The region labeled S is the collection of (Q, F) for which S is the equilibrium outcome, and similarly for D, M_α, and M_β.

Before analyzing the different regions in more detail, it is useful to state the basic intuition underlying the diagram. If the equilibrium involves only diversified firms operating, then two facts must obtain. Since all diversified firms are identical, their relative

outputs, $\tilde{q}^*_\alpha/\tilde{q}^*_\beta$, necessarily equal relative demand evaluated at the equilibrium prices, Q. Second, p^*_j must equal the marginal cost of good j in diversified firms. Thus, when demand is more skewed toward good j, $\tilde{q}^*_j/\tilde{q}^*_k$ must be greater, raising the marginal cost of q_j, implying a larger p^*_j.[8] Therefore, if a type j specialized firm could not profitably operate at some price p^*_j, it may be able to do so when demand is more skewed toward good j, involving a higher p^*_j. Returning to the diagram, and beginning at some point in D, raising Q (skewing demand toward good α) will eventually cause type α firms to operate alongside diversified firms (that is, (Q, F) is in M_α) unless the diversified firms' fixed cost (F) is appropriately low; that is, less than F_α. The higher the value of F, the less demand need be skewed to generate operation by type α firms. (Skewing demand toward good β lowers Q and can be treated symmetrically.) For large enough F, there is no pattern of demand for which diversified firms could be part of the equilibrium.

In brief then, the answer to the question of the conditions under which the various configurations will occur is as follows. When fixed costs for a diversified firm are large, S is the outcome irrespective of demand. Conversely, D invariably occurs if these fixed costs are sufficiently low. For intermediate values of the fixed cost, M_j is the outcome if demand is skewed toward good j, and D otherwise, with the degree of skewing required to generate M_j being smaller the larger is the diversified firms' fixed cost.

The interpretation of the dashed curve can now be given. At any point (Q, F) in the diagram, some marginal cost is implied for each good. Could these marginal costs equal equilibrium prices? The answer is in the affirmative if and only if relative demand at such prices is equal to the Q under consideration. The dashed line is the locus of such Q's, traced out (for a given set of product demands) as F varies. The figure is drawn for demands which are relatively "α-intensive." In the S region, since no diversified firms operate, the equilibrium Q cannot depend on F, and the dashed curve is vertical. As will be shown below, in the M_α region (again, M_β is symmetric), p^*_α is fixed and p^*_β changes in the same direction as F does. Thus when F falls, p^*_β declines along with it and equilibrium relative demand must involve lower Q. In the D region, changes in F imply changes in both p^*_j in the same direction, so equilibrium relative demand may either rise or fall. In the "base case," where relative demand depends only on relative prices, and equal proportional changes in $\tilde{q}_j$ leave relative marginal cost unaffected, the dashed line is vertical in region D as well as in region S.

E. *Some Details*

Analysis of the applications and extensions given below, as well as others the reader may wish to consider, is facilitated by some more detailed discussion of the different configurations.

First consider S equilibria. Under that configuration the relevant subset of (1′)–(3′) is

$$p_j - c'_j(q_j) = 0, \tag{1′a}$$

$$p_j q_j - F_j - c_j(q_j) = 0, \tag{2′a}$$

$$\text{and } D_j - N_j q_j = 0, \tag{3′a}$$

all for $j = \alpha, \beta$. The four conditions implied by profit maximization (1′a) and zero profit (2′a) determine the equilibrium prices $\bar{p}_j$ and firm-output levels $\bar{q}_j$ without reference to demand.[9] The equilibrium number of firms equates demand evaluated at $\bar{p}_j$ with total production. Notice that in the S region of Figure 1 prices $\bar{p}_j$ and quantities $\bar{q}_j$ always take on the same values, independent of both the structure of product demand and the fixed costs of the nonoperating diversified firms.

[8]This argument, and others to follow, makes use of the following consequence of the convexity of $c(\tilde{q}_\alpha, \tilde{q}_\beta)$: for given F and zero profits, an increase in $\tilde{q}_\alpha/\tilde{q}_\beta$ raises $\partial c/\partial \tilde{q}_\alpha$ and lowers $\partial c/\partial \tilde{q}_\beta$.

[9]Notice that whenever a type j firm operates in any equilibrium, $p_j = \bar{p}_j$ and $q_j = \bar{q}_j$ must hold.

Now suppose that at prices $\bar{p}_j$, a diversified firm is just able to earn zero profit.[10] Then the market-clearing conditions (3′) are

$$D_j - Nq_j^* - N_j\bar{q}_j = 0,$$

which gives two conditions to determine N, N_α, and N_β. The equilibrium values of these variables are thus not unique; a situation analogous to that in the single-good case in which two technologies yield identical values of minimum average cost. In the multiproduct model, when the diversified firms' fixed cost takes on the value labeled $\tilde{F}$ in Figure 1, the two technologies (that is, diversified production or pairs of specialized firms) available for producing any vector of outputs are equally efficient and may coexist in equilibrium. Raising F eliminates D while the outcome of lowering F depends on the pattern of demand. In Figure 1, a reduction in F causes the exit of type-β firms, as demand is skewed toward good α.

Turning to D equilibria, the relevant subset of (1′)–(3′) becomes

$$(1'\text{b}) \qquad p_j - \frac{\partial}{\partial \tilde{q}_j} c(\tilde{q}_\alpha, \tilde{q}_\beta) = 0,$$

$$(2'\text{b}) \qquad \sum_{j=\alpha,\beta} p_j \tilde{q}_j - F - c(\tilde{q}_\alpha, \tilde{q}_\beta) = 0,$$

$$(3'\text{b}) \quad \text{and} \qquad D_j - N\tilde{q}_j = 0.$$

There is just one zero-profit condition under Pure Diversification: (2′b). Along with the two profit-maximization conditions (1′b), there are only three restrictions to determine the four equilibrium values p_j^* and $\tilde{q}_j^*$. In contrast to the single output per firm case, there are many output vectors for the firm, which, if priced at marginal cost, yield zero profit. However, the two market-clearing conditions introduce only one additional variable, N, permitting (1′)–(3′) to characterize completely the equilibrium outcome. The intuition is simply that because there is but one type of firm in a D equilibrium, and each produces the same level of output of each good, those levels cannot be independent of demand. Indeed $\tilde{q}_\alpha^*/\tilde{q}_\beta^* = D_\alpha/D_\beta$ must hold. Also noteworthy is that for all (Q, F) pairs in the D region, p_j^*, equal to marginal cost in diversified firms, must be such that no type j firm could operate and earn nonnegative profits. That is, $p_j^* < \bar{p}_j$ for both j. Much more can be said about the manner in which the p_j^* vary in the D region. However, since the pattern of variation depends on the demand side, that discussion is deferred until demand shifts are considered in Section II.

Finally, the relevant part of (1′)–(3′) characterizing an M_j equilibrium, say M_α, is

$$(1'\text{c}) \qquad \begin{cases} p_\alpha - c'_\alpha(q_\alpha) = 0 \\ p_j - \dfrac{\partial}{\partial \tilde{q}_j} c(\tilde{q}_\alpha, \tilde{q}_\beta) = 0 \\ \qquad\qquad\qquad\qquad j = \alpha, \beta, \end{cases}$$

$$(2'\text{c}) \qquad \begin{cases} p_\alpha q_\alpha - F_\alpha - c_\alpha(q_\alpha) = 0, \\ \sum_{j=\alpha,\beta} p_j \tilde{q}_j - F - c(\tilde{q}_\alpha, \tilde{q}_\beta) = 0 \end{cases}$$

and

$$(3'\text{c}) \qquad \begin{cases} D_\alpha - N\tilde{q}_\alpha - N_\alpha q_\alpha = 0, \\ D_\beta - N\tilde{q}_\beta = 0 \end{cases}$$

The conditions (1′c) and (2′c) comprise five restrictions on the five variables p_j, $\tilde{q}_j$, and q_α. Thus, equilibrium p_j^* and firm-level outputs $\tilde{q}_j^*$ and q_α^* are again determined independently of demand. In particular, since type α firms operate, their profit-maximization and zero-profit conditions alone imply $p_\alpha^* = \bar{p}_\alpha$ and $q_\alpha^* = \bar{q}_\alpha$. Diversified firms' outputs $\tilde{q}_j^*$, and p_β^*, are then determined from

[10] That is, when

$$\bar{p}_j - \frac{\partial}{\partial \tilde{q}_j} c(\tilde{q}_\alpha, \tilde{q}_\beta) = 0$$

are solved for some $\tilde{q}_j$, then

$$\sum_{j=\alpha,\beta} \bar{p}_j \tilde{q}_j - F - c(\tilde{q}_\alpha, \tilde{q}_\beta) = 0$$

also holds for these same $\tilde{q}_j$.

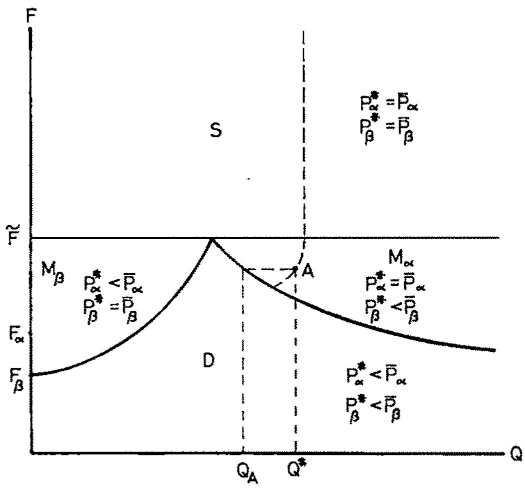


FIGURE 2

the remaining conditions in (1′c) and (2′c). Since no type-β firms can operate profitably, $p_\beta^* \le \bar{p}_\beta$. The derivation of the curve separating the M_α and D regions can now be given. At the intersection of this curve and the dashed line in Figure 1, the equilibrium values solve (1′c)–(3′c) above with $N_\alpha^* = 0$, and $p_\alpha^* = \bar{p}_\alpha$. That is, Pure Diversification and (trivially) Mixed Production give the same outcome. The associated q_α^*/q_β^* ratio is exactly that given by $D_\alpha(\bar{p}_\alpha, p_\beta^*)/D_\beta(\bar{p}_\alpha, p_\beta^*)$. For higher values of F—moving up the dashed line—$p_\alpha^* = \bar{p}_\alpha$ must still hold, as explained above. But under (6), the general expansion of scale necessary to maintain zero profits for diversified firms as F rises would raise the marginal cost of producing good α above $\bar{p}_\alpha$ if $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ did not change. Thus $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ necessarily falls, and N_α^* (>0) type α firms enter, to maintain market clearing. The curve separating the D and M_α regions is precisely the required $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ ratio associated with equilibrium on the dashed line, for any given F. Moreover, as F is raised further, $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ continues to fall, which raises the marginal cost of producing good β, hence increasing p_β^*. The ratio $\tilde{Q}$ in Figure 1 is the critical $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ at which F is so high that $p_\beta^* = \bar{p}_\beta$ occurs. Further increases in F, beyond $\tilde{F}$, imply that diversified firms no longer operate: an S equilibrium. The M_β region can be treated symmetrically.

Figure 2 summarizes. In particular, an equilibrium at the point labeled A yields

$$\tilde{q}_\alpha^*/\tilde{q}_\beta^* = Q_A, \quad \text{and}$$

$$D_\alpha(p_\alpha^*, p_\beta^*)/D_\beta(p_\alpha^*, p_\beta^*) = Q^*.$$

II. Predictions and Elaborations

What might be learned from this framework? The approach taken again parallels the standard analysis, but naturally focuses on the additional information obtained through extension to the multiple-output setting; indeed, S equilibria will not be discussed further. The descriptive features of the model are considered first. That is, given the setup above, what basic features should an investigator observe? Next, comparative static experiments suggested by the model are explained, although brevity dictates that most of the feasible experiments be left to the reader.[11] The selection offered here illustrates the manner in which the model may be manipulated, and highlights the experiments particular to the multiple-output environment. Finally, due to the general nature of the basic model, many applications will involve placing additional structure on its basic ingredients: demand, and fixed and variable costs. Simple examples of the type of results available from such extensions are provided.[12]

A. *Descriptive Results*

For comparative purposes, recall the basic description of equilibrium in the single-product setting. All firms produce at the same rate of output, equal to that which minimizes average cost, the value of which is price. Demand affects neither firm output level nor price, but determines the number of firms. That firms operate at "minimum efficient scale," and that demand shifts are met by

[11] More experiments are also available in the authors' 1986 paper.

[12] To avoid duplication, "Mixed Equilibrium" will henceforth refer to M_α. Exceptions will be obvious from the context.

entry and exit alone, are arguably the theory's key descriptive claims.

A D equilibrium also involves all firms producing in the same manner, but none of product prices, firm output levels, and the number of firms can be determined without reference to demand.[13] In an M_j equilibrium, the independence from demand is implied, but firms in a given industry do not all produce in the same manner.[14] Thus a new descriptive result, central to the multiproduct setting, is that whenever there are multiproduct firms, an investigator will not simultaneously observe both homogeneous actions by firms and independence of outputs and prices from demand within a single industry; a sharp contrast to the single-product result.

The second descriptive result is that when diversified firms operate, there can be at most one type of specialized firm.

Situations which these two central predictions help to organize are readily available. As an example using the first result, consider muffler and brake shops. There do not appear to be any firms specialized in either product, so suppose a D equilibrium is being observed. Given this assumption, it would be expected that such firms are very similar to one another, as they seem to be, and that demand plays a role in determining the muffler/brakes output ratio; that is, in climates where salt is used to melt snow, these firms sell more mufflers. Obvious as the latter might seem, it is not what the single-product analysis implies.

B. *Comparative Statics*

The impact of very general changes in relative demand may be determined directly from Figure 1. For example, consider an increase in $D_\alpha(p_\alpha, p_\beta)$, given any prices. This change causes the dashed curve in Figure 1 to shift to the right, for each F. Suppose the initial and subsequent equilibria are Purely Diversified. Then the demand shift raises $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ and p_α^*, and lowers p_β^*. Whether N^* rises or falls depends on the scale characteristics of $c(\tilde{q}_\alpha, \tilde{q}_\beta)$ and the demand elasticities. The demand shift may cause the equilibrium to become M_α, and in this case the predictions are unchanged except that entry of type α firms is also expected. In contrast, if the initial equilibrium was M_α, it remains so after the demand shift, and the sole impact is entry of type α firms.

This body of "demand shift" results may be aggregated to the third central prediction of the theory, which has a decidedly Ricardian flavor. Beginning in a D equilibrium, and shifting demand toward good j causes the price of good j to rise—much like an upward-sloping, long-run supply curve for good j. Production in diversified firms skews toward good j. But after a point, further demand shifts are met by entry of specialized firms, with relative output of diversified firms remaining at its most extreme value and price rising no further: an infinitely elastic long-run supply curve.

As an example, consider the production of printed news and advertising. Publications offering both are legion, and some offering ads alone are far from uncommon; fliers, catalogs, and so forth. There do not appear to be any pure news publications, in agreement with the second descriptive result. Thus, assume an M_j equilibrium. When the demand for ads rises, as during the pre-Christmas season for example, newspapers and magazines, while perhaps changing the nature of the ads, do not appear to devote more space to them, or to increase in number. But the number of fliers and catalogs increases dramatically. In contrast, when the demand for news rises, as it does during "national crisis," such as an assassination, "special editions," with the usual complement of ads, proliferate. Whether the number of fliers and catalogs is adversely affected, as theory would imply, is not immediately apparent. However, the independence of product prices from these demand shifts does seem to hold.

The effect of changes in fixed costs may also be analyzed. For simplicity, focus on

[13] Also, for $\tilde{q}_\alpha/\tilde{q}_\beta$ fixed at $\tilde{q}_\alpha^*/\tilde{q}_\beta^* = \mu$ production occurs at the minimum (with respect to $\tilde{q}_\beta$) of the analogue of average cost, $[F + c(\mu\tilde{q}_\beta, \tilde{q}_\beta)] \div \mu\tilde{q}_\beta$.

[14] This statement holds unless a "knife-edge" restriction on the cost functions is met.

the "base case" in which relative demand depends on relative prices alone, and relative marginal cost in diversified production is a function only of $\tilde{q}_\alpha/\tilde{q}_\beta$. Then the dashed line in Figure 1 is vertical in both the D and S regions. Now assume F increases, and that the initial and subsequent equilibrium is a D equilibrium. The new equilibrium, further up the dashed curve involves a simple expansion of scale, holding $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ constant, with p_α^* and p_β^* both rising by the same proportion. N^* necessarily falls, because demand is unchanged and each firm grows. If the subsequent equilibrium is M_α instead, $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ is reduced relative to its initial value, (recall $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ may be read off the curve separating D from M_α) while p_α^* and p_β^* again both rise. However, because entry of type α firms fixes p_α^* at $\bar{p}_\alpha$, the percentage increment to p_α^* is less than that to p_β^*, so that p_α^*/p_β^* falls. The cases in which the initial equilibrium is M_α can be handled in a similar fashion.

The impact of an increase in F_j is also straightforward, and an increment to F_α is depicted in Figure 3. Such a change implies a higher value for $\bar{p}_\alpha$ in the usual way, and the smallest relative demand at which type α firms may operate alongside diversified firms is therefore raised for each F; that is, the curve separating D from M_α shifts to the right. Turning to the dashed curve, it is unaffected in the original D region, because no type α firms were in operation for those parameter values even before the increase in F_α. However, in the S region, higher $\bar{p}_\alpha/\bar{p}_\beta$ implies reduced relative demand, shifting the dashed curve to the left, as also occurs in M_α. Given these changes, if the initial equilibrium was M_α, the new equilibrium may either remain an M_α or become D. If M_α remains, p_α^* rises to the new higher value of $\bar{p}_\alpha$, and diversified firms raise $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$, implying a decline in p_β^*. If the new equilibrium is D, the predictions are the same except that p_α^* rises to some value less than $\bar{p}_\alpha$, accompanying exit of type α firms.[15]

[15]Also, note that if the initial equilibrium was S, it may be M_α subsequently. The increase in F_α has led to the demise of type β firms. The reason is that diversified firms may operate at a higher $\tilde{q}_\alpha/\tilde{q}_\beta$ ratio when F_α is greater, lowering the marginal cost of good β in diversified firms, thus permitting their operation given (Q, F) pairs which would not have permitted that outcome initially.

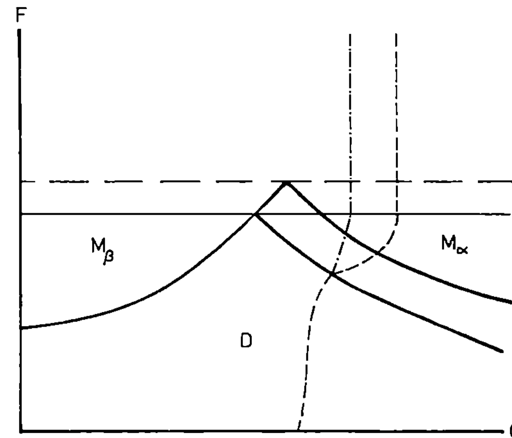


FIGURE 3

In what follows, several experiments may be cast in terms of one of the following two parameter changes: (i) an equal increase in all fixed costs; or (ii) an increase in F and F_j, with F_k unchanged. It is thus appropriate to provide the analysis of these alterations in the model's parameters at this point.

The first case is displayed in Figure 4. Consider raising both F_j, with the increase in F ignored for the moment. As indicated above, the curve separating D from $M_\alpha(M_\beta)$ shifts to the right (left). Consequently, the D region expands, S contracts, and the M_j regions lose some area to D but gain from S.[16] Interest in this experiment derives from the fact that the vertical change in curves separating D from M_j exceeds the change in F_j (which equals the increase in F). Thus, an equal increase in all fixed costs fosters diversification in the sense that (i) if diversified firms were part of the equilibrium prior to the change, they must also be in operation following it; (ii) some equilibria in which diversified firms were excluded

[16]This result imposes the (sufficient) condition that $\partial c/\partial \tilde{q}_j$ is convex.

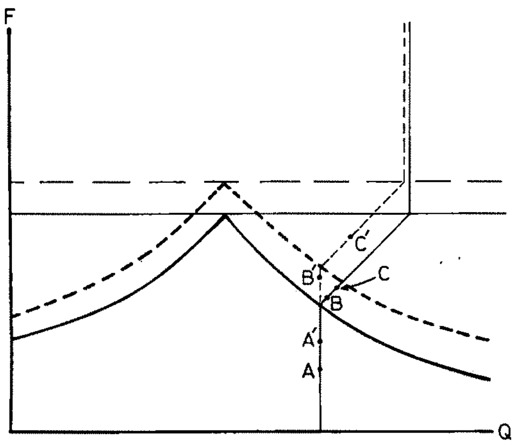


FIGURE 4

initially involve their operation after the change; and (*iii*) initially mixed equilibria may involve only diversified production after the change. Note however, that it is not always the case that specialized firms will exit in response to an equal increase in all fixed costs. Indeed, if relative demand is quite elastic, the equilibrium may be D initially and M_α subsequently.

How do prices and quantities respond to this change? If the initial and subsequent equilibria are D equilibria, the earlier analysis of a change in F alone again applies ($A \to A'$ in the figure). If the initial equilibrium was M_α, irrespective of whether the new equilibrium is D ($B \to B'$) or still M_α ($C \to C'$), $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$ must rise. Together with the effect of the diversified firms' general expansion of scale, an increase in p_α^* is necessarily implied, while p_β^* may either rise or fall.

The second experiment, which alters F and F_j, F_k remaining unchanged, may be analyzed similarly. Doing so yields the same conclusions as a change in all fixed costs so long as the increase in F is not so large as to eliminate diversified firms entirely, as might occur if the initial equilibrium were close to the curve separating D from M_k (that is, this curve does not shift when F and F_j only are increased).

C. *Elaborations*

Demand. The impact of general changes in demand was analyzed above. However, a noteworthy special case involves home production. Assume all consumers are identical and have CES preferences defined on two home commodities z_1 and z_2:

$$U = \left[\gamma z_1^{-\delta} + (1-\gamma) z_2^{-\delta}\right]^{-1/\delta},$$

where $\gamma \in (0,1)$ and $\delta \in (-1,\infty)$ are fixed parameters. Also, assume z_1 and z_2 are produced according to the home technologies

$$z_1 = A x_\alpha^a t_1^{1-a},$$

and

$$z_2 = B x_\beta^b t_2^{1-b},$$

where x_j is the quantity of good j used in home production, a and b are parameters, and t_1 and t_2 are times allocated to such production. Given a wage rate of w, relative demand implied by this setup is[17]

$$\left(\frac{\gamma}{1-\gamma}\right)^\sigma \left[\frac{A}{B}\left(\frac{b}{1-b}\right)^{1-b}\left(\frac{1-a}{a}\right)^{1-a}\right]^{\sigma-1} \times p_\alpha^{a(1-\sigma)-1} p_\beta^{b(\sigma-1)+1} w^{(b-a)(1-\sigma)},$$

where $\sigma = 1/(1+\delta)$. Changes in the parameters of preferences (γ, σ) or home technologies (A, B, a, b), or the prices of other inputs to home production (w) all have effects which may be analyzed in the manner illustrated in the previous subsection.

Fixed Costs. Fixed costs can be given a richer structure in numerous ways. One interesting approach is to suppose that production of good j requires a fixed amount, $\bar{V}_j$, of a good j-specific factor V_j that may be thought of as the organizational requirements for the design, production, and marketing of good j. $V_j \geq \bar{V}_j$ is a constraint

[17]This analysis assumes that not all available time is used in home production. The opposite assumption may be handled similarly.

irrespective of whether production of good j takes place in a type j or diversified firm, but V_j may be obtained in a variety of ways

$$V_j = g^j(m, s_j).$$

Here $g^j(\cdot)$ describes the technology (with the standard neoclassical properties) through which two inputs (m and s_j) are combined to produce V_j. Input m (for "management," but subsuming accounting, etc.) operates as a pure public input in the sense that its activities provide a service flow to all other processes within the firm. Input s_j (for "supervision," but including product design, marketing...) also has a public input character, but only with respect to activities that yield good j as output. Prices are denoted r and w_j, for m and s_j, respectively.

To satisfy the constraint $V_j \geq \overline{V}_j$ at least cost, a type j firm solves

$$\min_{m, s_j} rm_j + w_j s_j$$

$$\text{subject to } g^j(m_j, s_j) \geq \overline{V}_j,$$

where m_j is the quantity of m used by a type j firm. In contrast, a diversified firm solves[18]

$$\min_{\tilde{m}, \tilde{s}_\alpha, \tilde{s}_\beta} r\tilde{m} + \sum_j w_j \tilde{s}_j$$

$$\text{subject to } g^j(\tilde{m}, \tilde{s}_j) \geq \overline{V}_j, \qquad \text{for } j = \alpha, \beta.$$

Using asterisks to distinguish the cost-minimizing choices, implied values for the fixed costs are

$$F = r\tilde{m}^* + \sum_j w_j \tilde{s}_j^*,$$

and

$$F_j = rm_j^* + w_j s_j^*.$$

[18]In keeping with the convention used throughout, input quantities used by diversified firms will be distinguished by a tilde.

Two immediate implications in an M_j equilibrium are $\tilde{m}^* > m_j^*$ and $\tilde{s}_j^* < s_j^*$. That is, diversified firms will make use of more management and less good j-specific supervision.

Now, note that increases in $\overline{V}_j$ or w_j leave F_k constant, but raise both F and F_j, and the increase in F_j can be shown to exceed the rise in F. From the results of the previous subsection then, if the initial equilibrium is M_α, an increase in $\overline{V}_\alpha$ or w_α fosters diversification, and raises both p_α^* and $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$. (Refer to Figure 4 again.) Increments to r, on the other hand, raise F by more than F_j. Clearly diversification may not be encouraged, but, again focusing on the M_α equilibrium, p_α^* rises along with $\tilde{q}_\alpha^*/\tilde{q}_\beta^*$.

Variable Costs. Structure can be placed on variable costs in many ways. One simple specification is to suppose

$$c(\tilde{q}_\alpha, \tilde{q}_\beta) = \sum \rho_j \eta_j \tilde{q}_j + \tilde{c}(\tilde{q}_\alpha, \tilde{q}_\beta),$$

where ρ_j is the price of a good j-specific variable input, say y_j, η_j is the exogenous factor/output ratio for y_j, $\tilde{c}(\tilde{q}_\alpha, \tilde{q}_\beta)$ is the cost of all other variable factors. From (7), $c_\alpha(q_\alpha)$ is simply $c(\tilde{q}_\alpha, 0)$ for $q_\alpha = \tilde{q}_\alpha$, etc. This version of $c(\cdot)$ would be appropriate if y_j were a material input not used in the production of k.

Under this formulation, an increase in ρ_j or η_j raises the marginal cost of good j at any given output for both diversified and specialized firms, the marginal cost of k being unaffected. Also, the change in the marginal cost of j does not depend on $\tilde{q}_j$ or q; thus, the partition of the (Q, F) plane in the above figures is unaltered. The sole impact of this change is thus in terms of the position of the dashed curve. In Figure 1, for example, an increment to ρ_α or η_α raises the relative marginal cost of good α, in which case the dashed curve shifts to the left. If the initial and subsequent equilibria are in M_α, the behavior and number of diversified firms is unaffected by the change, with the reduction in quantity demanded due to the higher-product price being met by exit of

type α firms. The other cases may be handled similarly.[19]

III. Summary

The purpose of this paper was to present a framework which would allow analysis of a variety of situations involving multiproduct firms to proceed with the same ease, rigor, and clarity available in the conventional single-product setting. This goal was accomplished by developing a two-good model structured in a fashion very similar to the single-product model. Consumers' behavior was summarized by market demand for each good. Firms could specialize in the production of one good, or adopt a technology yielding both. The alternative technologies were allowed to differ in terms of fixed and variable costs. It was shown that there are three possible types of equilibrium configurations: Pure Specialization, Pure Diversification, and Mixed Production, corresponding to only specialized firms operating, only diversified firms operating, or production by a combination of diversified firms and exactly one type of specialized firm. The precise outcome depends on the model's exogenous entities (demand, fixed, and variable costs). A simple diagram summarized this dependence.

Section II then showed that the model is useful by (i) presenting the basic descriptive aspects of the economy; (ii) deriving the effect of changes in the underlying exogenous factors on the equilibrium configuration of product prices and outputs; and (iii) providing a collection of extensions/applications.

[19] More results may be obtained by placing structure on the production technology underlying the cost functions. For example, an appealing notion is that specialized firms may utilize production processes more clearly suited to the production of a particular good (for example, specialized machinery, etc.), but may not exploit the public good aspects of centralized management in the same way diversified firms can. In the authors' 1986 paper, a detailed model of this situation was presented. Therein diversified firms used factor proportions intermediate to those that would be chosen by specialized firms (when the choice made by the latter differ by firm type). It was shown that when technologies for producing different goods are more similar (in a well-defined fashion) diversification is fostered. The model thus helps to organize observations such as that automobiles and trucks are often manufactured in the same plant, and why bicycles and sewing machines are not, though doing so was common earlier in this century.

REFERENCES

Allen, Roy G. D., *Mathematical Analysis for Economists*, London: Macmillan, 1938.

Bailey, Elizabeth, E. and Friedlander, Ann F., "Market Structure and Multiproduct Industries," *Journal of Economic Literature*, September 1982, *20*, 1024–48.

Baumol, William J., Panzar, John C. and Willig, Robert D., *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich, 1982.

Hicks, John R., *Value and Capital*, London: Oxford, 1939.

Laitinen, Keith, *A Theory of the Multi-Product Firm*, New York: North-Holland, 1980.

MacDonald, Glenn M. and Slivinski, Alan, "A Positive Analysis of Multiproduct Firms in Market Equilibrium," mimeo., 1986.

Marshall, Alfred, *Principles of Economics*, London: Macmillan, 1920.

Samuelson, Paul A., *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.

Economic Organization with Limited Communication

By ROBERT M. TOWNSEND*

This paper presents formal, stylized representations of communication-accounting systems: oral assignment, portable object, written message, and telecommunication systems are considered. The environments that allow this formalization are characterized by spatial separation, private information, and a need to keep track of past actions, transfers, and shocks.

In environments that have spatial separation and private information, beneficial multilateral arrangements can depend critically on agents' ability to communicate to one another values of contemporary shocks and to keep track of histories of past transfers or past-announced shocks. This paper formalizes this idea and focuses on communication-accounting systems. The theory of this paper allows a formal, stylized representation of a variety of systems and allows one to make precise the sense in which various systems are more or less limited. Oral assignment systems, portable object systems, written message systems, and telecommunication systems are considered.

In its method this paper follows the literature on contract theory and mechanism design, of Milton Harris and Robert Townsend (1981), Roger Myerson (1979), and Townsend (1982), for example, stressing private information and incentives. The idea, essentially, is to specify the agents' endowments and preferences and the production technology available to them, and to be precise about the information structure. Then, rather than imposing a fixed-contract form or fixed-resource allocation scheme, one considers a broad class of arrangements and determines the constraints implied by private information. One then goes on to determine Pareto-optimal arrangements by maximizing weighted sums of the agents' utilities, subject to the obvious resource constraints and these derived, information, incentive compatibility constraints.

This paper takes this method one step further by making explicit both the agents' locations at various times and the technology of communication available to agents over space and over time. Exogenous variations in the technology of communication thus cause endogenous variations in the derived incentive constraints and, in this way, in the context of the class of maximization problems, one can capture formally the idea that communication systems matter and that particular systems may be more or less limited. Indeed, oral assignment systems, portable object systems, written message systems, and electronic telecommunication systems can be ordered: these are successively less limited.

The communication-accounting systems considered in this paper are motivated by observations from "simpler," historical, and contemporary economies. Oral communication systems are those in which agents must literally get together with one another in order to give instructions to one another or to execute some prespecified arrangement. For example, banking at the beginning of the commercial revolution in Europe took the form of an oral assignment system, as described by Abbott Usher (1943). Both the

*Department of Economics, University of Chicago, Chicago, IL 60637. I gratefully acknowledge all helpful comments from the participants of workshops at Carnegie-Mellon University, the University of Chicago, Cornell University, the Federal Reserve Bank of Minneapolis, the University of Iowa, the University of Pennsylvania, the University of Rochester, Stanford University, the University of Virginia, the University of Western Ontario, the University of Wisconsin, and students in Economics 339 at the University of Chicago; the excellent research and computational assistance from Christopher Phelan; and financial research support from the Alfred P. Sloan Foundation and the National Science Foundation grant nos. SES-8406771 and SES-8708242.

need not be communicated because, by assumption, such states of the world are observed by everyone and have been incorporated into the prespecified agreement to allocate labor to production, to transfer goods, and to reallocate agents over space. To allow for communication then, that is, to allow for a discussion of communication-accounting systems, something must be done to alter the model.

The alteration discussed at length in this paper is the incorporation of private information. That is, in addition to uncertainty, realizations of some of the random variables of the model are presumed to be seen by subsets of agents only. Indeed, this is a natural way to allow for communication, because it is known from the work of Myerson (1979) and of Harris and Townsend (1981) that programs for the determination of Pareto-optimal allocations can be utilized if agents are given the opportunity to announce (communicate) values for the subset of random variables, the realizations of which they alone see. That is, one can impose without loss of generality incentive compatibility constraints which ensure that these privately observed realizations are announced truthfully.

Because this idea is at the heart of the analysis, here it is best to review it in the context of a relatively simple environment, one without any essential dynamics and without any spatial separation. Further, to avoid complications introduced by putting private information on quantities, something which will be contemplated later, it is best to begin by supposing privately observed and unverifiable shocks to preferences. The motive for trade in the planning period is insurance.

Thus imagine an economy with just two agents, named a and b, and two time periods, a planning date and a consumption date. Agent a has an endowment e^a in the consumption date, a nonnegative vector of goods. Typically, we shall consider cases in which there is only one good, or two, but the dimension of e, and of consumptions below, can be arbitrary if finite. Agent a has preferences in the consumption date over consumption vector c^a as represented by utility function $U^a(c^a,\theta^a)$. Here θ^a is a shock to a's preferences, observed by agent a alone at the beginning of the consumption date. Also, θ^a takes on values in some set Θ^a. Typically, set Θ^a contains two values, or three, and each value is at most two-dimensional, but again the number of values and the dimension can be arbitrary if finite. Agent b has endowment e^b and preferences $U^b(c^b)$ in the consumption period. Utility functions are concave, and there is no production technology. Further, this structure is common knowledge. That is, everything is known up to shocks θ^a, presumed as of the planning period to occur with probabilities $p(\theta^a)$.

In the planning period both agents commit themselves to an allocation of the consumption goods in the consumption period contingent on agent a's announcement in the consumption period of the shocks he has experienced. That is, both agents precommit to pool their endowments and redistribute the total under a prespecified plan. Further, as a technical matter, this shock contingent allocation rule can be a lottery, $\pi(c|\theta^a)$, specifying the probability of consumption bundle $c=(c^a,c^b)$ to agents a and b. This will ensure that the programming problem is concave (actually linear), despite private information, though no essential use will be made of the lotteries in the solutions reported below. Thus letting $C=\{c=(c^a,c^b); c^a+c^b\le e^a+e^b\}$ denote the space of feasible consumptions, presumed for simplicity to be finite as if there were some indivisibility, the programming problem for the determination of Pareto-optimal allocations is

Program 1: Maximize by choice of the $\pi(c|\theta^a)$ the objective function

$$(1)\quad \lambda^a\Big\{\sum_{\theta^a}p(\theta^a)\sum_c\pi(c|\theta^a)U^a(c^a,\theta^a)\Big\} + \lambda^b\Big\{\sum_{\theta^a}p(\theta^a)\sum_c\pi(c|\theta^a)U^b(c^b)\Big\},$$

subject to the incentive compatibility constraints,

$$(2)\quad \sum_c\pi(c|\theta^a)U^a(c^a,\theta^a) \ge \sum_c\pi(c|\tilde{\theta}^a)U^a(c^a,\theta^a)\quad \forall\theta^a,\tilde{\theta}^a\in\Theta^a.$$

One can easily append onto this program *ex ante* participation constraints,

$$(3)\quad \sum_{\theta} p(\theta^a)\sum_{c}\pi(c|\theta^a)U^a(c^a,\theta^a) \geq \sum_{\theta} p(\theta)U^a(e^a,\theta^a),$$

$$(4)\quad \sum_{\theta} p(\theta^a)\sum_{c}\pi(c|\theta^a)U^b(c^b) \geq U^b(e^b).$$

In fact, weights λ^a and λ^b will be chosen implicitly in the examples below by the presumption that the expected utility of agent a is maximized subject to constraint (4) for agent b.

Solutions to program 1 can be generated as solutions to linear programs, computed numerically. Two examples provide illustrations to be carried through the subsequent analysis. The first has only one consumption good; the second, two.

For the first example, the utility function of agent a is of the form

$$U^a(c,\theta^a) = (c)^{\theta^a},$$

so that agent a is risk averse for each parameter θ^a, with $\theta^a \in (.4,.5,.6)$, each possibility occurring with probability one-third. The utility function of agent b is of the form $U^b(c) = c$, so that b is risk neutral. Also let $e^a = e^b = 5$. Then, ignoring the discrete use of the consumption space, the full information solution would allocate the consumption good in such a way as to equate weighted marginal utilities over states. That is, agent a would receive the consumption good when he is urgent, at least at $\theta^a = .6$, and give it up when he is patient, at least at $\theta^a = .4$. On the other hand, with private information, the incentive constraints do not allow this dependence, and the solution degenerates to autarky (despite allowance for lotteries). This is natural since there is only one good and more is preferred to less. In this case, then, private information is quite damning to the social arrangement. Though trivial, the full information and private information solutions are reported in Table 1.

For the second example there are two goods, denoted generically as x and y, and θ is two-dimensional. The utility function of agent a is of the form

$$U^a(c_x,c_y,\theta^a) = (c_x)^{\theta^a_x} + (c_y)^{\theta^a_y},$$

with $(\theta^a_x,\theta^a_y) \in \{(.4,.6),(.6,.4)\}$, each possibility occurring with probability one-half. Agent b continues to be linear, now in both goods. Endowments are $e^a_x = e^b_x = e^a_y = e^b_y = 5$. The full information and private information solutions are reported in Table 2.

TABLE 1—SINGLE-PERIOD SOLUTION, ONE GOOD

Values for θ^a	c^a (Full Information)	c^a (Private Information)
.4	2.2	5
.5	4	5
.6	8.8	5

TABLE 2—SINGLE-PERIOD SOLUTION, TWO GOODS

Values for θ^a_x,θ^a_y	(Full Information) (c^a_x,c^a_y)	(Private Information) (c^a_x,c^a_y)
(.4,.6)	(2,8)	(2,8)
(.6,.4)	(8,2)	(8,2)

The full information and private information solutions are identical, because the incentive constraints (2) are not binding; that is, with two commodities, goods can be assigned optimally in such a way that agent a self-selects. (There is some *ex ante* beneficial insurance associated with the solution.)

Again, there is presumed to be perfect, costless enforcement of a selected, *ex ante* optimal arrangement. It seems best to take this as given, if only as an approximation, without spelling out the mechanics of the enforcement procedure. Otherwise, without any enforcement mechanism, one must suppose full commitment. That is, agents a and b must show up at the consumption date; agents a and b must pool their endowments then as if these had to be placed on preset conveyors, activated by weight; agent a is restricted to announce one value for θ^a in the set Θ^a, as if a computer were pro-

grammed to accept one of a prespecified set of values; consumption is reallocated on the conveyors in accordance with the announced value θ^a and the preset computer program; and finally agents eat in private, without redistribution. Such a technology may read more as science fiction than as realistic economics, but it is important to note that in principle there could exist a transfer function, message space technology that would allow one to implement a solution to a socially optimal program. That is, private information is the only impediment to trade.

II. Multiperiod Arrangements and the Gain from Intertemporal Links

The analysis above is easily extended to accommodate some nontrivial dynamics. In particular, suppose agents a and b remained paired with each other over two consumption dates t, $t=1,2$. Let preference shocks be experienced by agent a at the beginning of date t. That is, let shock θ_1^a enter into the utility function of agent a at date $t=1$ and let shock θ_2^a enter into the utility function of agent a at date $t=2$. However, some persistence in shocks is allowed. That is, $\theta_2^a = f(\theta_1^a, \delta)$, where f is a deterministic function and δ is a nondegenerate random variable, possibly dependent on θ_1^a. This allows as a special case no direct intertemporal links in shocks apart from Markov dependence in the probabilities, $p(\theta_2^a|\theta_1^a)$. Let e_t^a and e_t^b denote the vector of endowments of agents a and b at dates t, $t=1,2$, presumed for simplicity not to vary with date t. Here and below let β denote the common discount rate for time-separable utilities.[1] Finally, let $\pi_1(c|\theta_1^a)$ denote the allocation rule at date 1 and $\pi_2(c|\theta_1^a,\theta_2^a)$ denote the allocation rule at date 2, both of these specifying probabilities on set $C=\{c=(c^a,c^b), c^a+c^b\le e^a+e^b\}$ as before, still supposing no storage. Then the program for the determination of a private information Pareto-optimal arrangement is

Program 2: Maximize by choice of $\pi_1(c|\theta_1^a)$ and $\pi_2(c|\theta_1^a,\theta_2^a)$ the objective function

$$\begin{aligned}(5)\quad &\lambda^a\Big\{\sum_{\theta_1^a} p(\theta_1^a)\sum_c \pi_1(c|\theta_1^a)U^a(c^a,\theta_1^a)\\ &+\beta\sum_{\theta_1^a} p(\theta_1^a)\sum_{\theta_2^a} p(\theta_2^a|\theta_1^a)\\ &\times\sum_c \pi_2(c|\theta_1^a,\theta_2^a)U^a(c^a,\theta_2^a)\Big\}\\ &+\lambda^b\Big\{\sum_{\theta_1^a} p(\theta_1^a)\sum_c \pi_1(c|\theta_1^a)U^b(c^b)\\ &+\beta\sum_{\theta_1^a} p(\theta_1^a)\sum_{\theta_2^a} p(\theta_2^a|\theta_1^a)\\ &\times\sum_c \pi_2(c|\theta_1^a,\theta_2^a)U^b(c^b)\Big\},\end{aligned}$$

subject to incentive constraints at date $t=2$, for all $\tilde\theta_1^a$ announcements in the past, for all actual contemporary values θ_2^a, and for all counterfactual announcements $\tilde\theta_2^a$,

$$\begin{aligned}(6)\quad &\sum_c \pi_2(c|\tilde\theta_1^a,\theta_2^a)U^a(c^a,\theta_2^a)\\ &\ge\sum_c \pi_2(c|\tilde\theta_1^a,\tilde\theta_2^a)U^a(c^a,\theta_2^a),\end{aligned}$$

and subject to incentive constraints at date $t=1$ for θ_1^a actuals and $\tilde\theta_1^a$ counterfactuals,

$$\begin{aligned}(7)\quad &\sum_c U^a(c^a,\theta_1^a)\pi_1(c|\theta_1^a)+\beta\sum_{\theta_2^a} p(\theta_2^a|\theta_1^a)\\ &\times\sum_c U^a(c^a,\theta_2^a)\pi_2(c|\theta_1^a,\theta_2^a)\\ &\ge\sum_c U^a(c^a,\theta_1^a)\pi_1(c|\tilde\theta_1^a)\\ &+\beta\sum_{\theta_2^a} p(\theta_2^a|\theta_1^a)\\ &\times\sum_c U^a(c^a,\theta_2^a)\pi_2(c|\tilde\theta_1^a,\theta_2^a).\end{aligned}$$

[1]These solutions are computed for $\beta=.95$. Also the solution is reported as deterministic, but in fact the generated solution for a grid of 101 points between 0 and 10 displayed lotteries. For example at $\theta_1^a=.6$, the computed solution is 7.2 with Prob .4872 and 7.3 with Prob .5128. As this lottery is an artifact of grid size, and would disappear with a continuum of possible consumptions, the reported solution is delivered by linear interpolation. Similar interpolations are done for the tables that follow.

TABLE 3—MULTIPERIOD PRIVATE INFORMATION SOLUTION, ONE GOOD[a]

Values for θ_1^a	c^a at Date 1	Values for θ_2^a	c^a at Date 2
		.4	6.55
.4	3.1	.5	6.55
		.6	6.55
		.4	4.87
.5	5.1	.5	4.87
		.6	4.87
		.4	3.1
.6	7.25	.5	3.1
		.6	3.1

[a]See fn. 1.

TABLE 4—MULTIPERIOD PRIVATE AND FULL INFORMATION SOLUTION, TWO GOODS

Values for $(\theta_{1x}^a, \theta_{1y}^a)$	(c_x^a, c_y^a)	Values for (δ_x, δ_y)	Values for $\theta_{2x}^a, \theta_{2y}^a$	(c_x^a, c_y^a)	
		(1,1)	(.6, .4)	8.01	2.0
(.4,.6)	(2,8)	(.5, 1.5)	(3., .6)	1.0	8.0
		(1.5, .5)	(.9, .2)	10.0	0.82
		(1,1)	(.4, .6)	2.0	8.0
(.6,.4)	(8,2)	(.5, 1.5)	(.2, .9)	0.82	10.0
		(1.5, .5)	(.6, .3)	8.0	1.0

What bears emphasis in this program is the possible dependence of allocation rule at date $t = 2$ on announcement of parameter θ_1^a at date 1. Indeed, suppose there is only one good, so that beneficial trade under private information is difficult, as the no-insurance, autarkic, private information solution of Table 1 emphasizes. Suppose further, to bias the case against intertemporal tie-ins, that there is no functional persistence in the preference shocks and no Markov dependence in probabilities. That is, each shock is drawn with equal likelihood in each period. Still, as displayed in Table 3, for the environment of Table 1, there is some insurance at date $t = 1$ (but not at date $t = 2$) achieved by intertemporal dependence.

This dependence does not appear in the full information, full insurance solution, the full information solution of Table 1 reported twice, once for each period. In fact, this dependence would not appear in the full information solution even with nontrivial Markov probabilities, since the full information rule remains the same: allocate consumptions so as to equate weighted marginal utilities for every contemporary state. The private information solution displays dependence because intertemporal tie-ins are used to circumvent the damning effects of the incentive constraints; low consumption at date $t = 1$ is tied to high consumption at date $t = 2$, and conversely. Again, this allows for some insurance.

With two (or more) goods, the date $t = 1$ incentive constraints are not so damning. In fact, as is evident from Table 2, it is possible to construct examples without binding incentive constraints at date 1, and if there were no persistence in shocks, intertemporal tie-ins would not be needed. On the other hand, persistence in shocks can deliver intertemporal tie-ins. For example, suppose there are two goods at each date. The utility function of agent a at date 1 is of the form

$$U^a(c_x, c_y, \theta_1^a) = (c_x)^{\theta_{1x}^a} + (c_y)^{\theta a}_{1y}$$

with $\quad (\theta_{1x}^a, \theta_{1y}^a) \in \{(.4,.6),(.6,.4)\},$

each with equal probability, and at date 2 of the form

$$U^a(c_x, c_y, \theta_2^a) = (c_x)^{(1-\theta_{1x}^a)\delta_x} + (c_y)^{(1-\theta_{1y}^a)\delta_y}$$
$$= (c_x)^{\theta_{2x}^a} + (c_y)^{\theta_{2y}^a}$$

with

$$(\delta_x, \delta_y) = \begin{cases} (1,1) & \text{with Prob .96} \\ (.5,1.5) & \text{with Prob .02} \\ (1.5,.5) & \text{with Prob .02} \end{cases}$$

Table 4 reports the solution.[2] To be noted is that the family of allocations available for agent a at date 2 depends on the announced (and actual) parameter draw of agent a at date 1.

Regardless of how the tie-ins are generated, the private information optimal solu-

[2]Again endowments are five uniformly and $\beta = .9$.

tion is damaged if tie-ins are not allowed, that is, if there were no record of preference shock announcements of agent a at date $t=1$ (so that at most reannouncements are viable). For the one good example above, this is obvious; with no tie-ins there is no trade, and the solution is autarky for both periods. For the two-good example, intratemporal reallocations are still viable, as in the single-period solution of Table 2. But the absence of direct tie-ins would be associated with a loss of utility.[3]

It bears repetition that perfect and costless enforcement of the private information optimal mechanism is still assumed. That is, neither agent can walk away from the agreed-upon arrangement at the end of date 1, and both are committed to comply with the multiperiod message space and transfer function requirements.

III. Optimal Multilateral Arrangements with Spatial Separation but Full Communication

With this investment in mechanism design, one can now extend the two-period model to include four agents and two locations, as a base for discussion in the sections which follow of various communication-accounting systems. This section focuses on the benefit from quadrilateral, rather than bilateral, arrangements.

The four-agent, two-location, two-period model is like the two-agent (one location), two-period model of Section II except that there are two agents of type a, named a and a', and two agents of type b, named b and b'. Agent a is presumed to stay at location 1 over the two periods of his lifetime, and agent a' stays at location 2. But agents b and b' switch locations between dates one and two, in accordance with Table 5.

Agent a experiences privately observed preference shocks at the beginning of each date t, $t=1,2$, and similarly for agent a', and, until otherwise specified, the distribution determining shocks for agent a is independent of the distribution determining shocks for agent a'. Agents b and b' experience no shocks. The notation for endowments is as before, and for simplicity these do not vary over agents or over time.

With $\pi_{it}(\cdot)$ as general notation for the allocation rule for the two agents present at location i and date t, $i=1,2$, $t=1,2$, let $\pi_{11}(c|\theta_1^a)$ and $\pi_{12}(c|\theta_1^a,\theta_2^a)$ denote the probabilities of consumption bundle c in common space C in location 1 at dates 1 and 2, respectively, conditioned on the announcements of agent a, and allowing intertemporal tie-ins. Also let $\pi_{21}(c|\theta_1^{a'})$ and $\pi_{22}(c|\theta_1^{a'},\theta_{2}^{a'})$ denote the corresponding probabilities of consumption bundle c at location 2, conditioned on the announcements of agent a'. Then the program for the determination of private information, full communication, Pareto-optimal allocations is:

Program 3: Maximize by choice of the $\pi_{it}(\cdot)$ the objective function

$$
\begin{aligned}
(8)\quad & \lambda^a\Big\{\sum_{\theta_1^a} p(\theta_1^a)\sum_c \pi_{11}(c|\theta_1^a)U^a(c^a,\theta_1^a) \\
&+\beta\sum_{\theta_1^a} p(\theta_1^a)\sum_{\theta_2^a} p(\theta_2^a|\theta_1^a)\sum_c \pi_{12}(c|\theta_1^a,\theta_2^a)U^a(c^a,\theta_2^a)\Big\} \\
&+\lambda^b\Big\{\sum_{\theta_1^a} p(\theta_1^a)\sum_c \pi_{11}(c|\theta_1^a)U^b(c^b) \\
&+\beta\sum_{\theta_1^{a'}} p(\theta_1^{a'})\sum_{\theta_2^{a'}} p(\theta_2^{a'}|\theta_1^{a'})\sum_c \pi_{22}(c|\theta_1^{a'},\theta_2^{a'})U^b(c^b)\Big\} \\
&+\lambda^{a'}\Big\{\sum_{\theta_1^{a'}} p(\theta_1^{a'})\sum_c \pi_{21}(c|\theta_1^{a'})U^{a'}(c^{a'},\theta_1^{a'}) \\
&+\beta\sum_{\theta_1^{a'}} p(\theta_1^{a'})\sum_{\theta_2^{a'}} p(\theta_2^{a'}|\theta_1^{a'})\sum_c \pi_{22}(c|\theta_1^{a'},\theta_2^{a'})U^{a'}(c^{a'},\theta_2^{a'})\Big\} \\
&+\lambda^{b'}\Big\{\sum_{\theta_1^{a'}} p(\theta_1^{a'})\sum_c \pi_{21}(c|\theta_1^{a'})U^{b'}(c^{b'}) \\
&+\beta\sum_{\theta_1^a} p(\theta_1^a)\sum_{\theta_2^a} p(\theta_2^a|\theta_1^a)\sum_c \pi_{12}(c|\theta_1^a,\theta_2^a)U^{b'}(c^{b'})\Big\},
\end{aligned}
$$

[3]In the absence of direct tie-ins, that is, in the absence of a record of past announcements, agent a might be required to reannounce date 1 preference shocks and to announce a value for contemporary shocks. In this scheme, reannouncements of date 1 shocks may matter, and in that weak sense there are intertemporal tie-ins. But the solution is worse than if there were a record of past announcements, of $\tilde{\theta}_1^a$ in program 2.

TABLE 5—AGENT PAIRINGS IN THE FOUR-AGENT, TWO-LOCATION MODEL

	Location	1	2
Date	1	(a,b)	(a',b')
	2	(a,b')	(a',b)

subject to incentive constraints for agent a at date $t=1,2$, identical to incentive constraints (6) and (7) above, except with $\pi_{1t}(\cdot)$ replacing π_t, and subject to similar incentive constraints for agent a' at date $t=1,2$.

Some results are already implicit in this program but deserve elaboration. First, the full information solution to the program would still equate weighted marginal utilities of the two agents present at each location by appropriate distribution of the total endowment of the two agents present. As only the contemporary preference shock of agent a would enter as a genuine variable into these equations for location 1, and that of agent a' for location 2, full information optimal rules depend at most on these parameters. In particular, the shock experienced by agent a' at date 1 and location 2 has no bearing on the transfer at date 2 and location 1 even though agent b' is present at both locations. That is, the history experienced by b' would not matter. This argument carries over to the private information, full communication program in question, as can be deduced by a study of first-order conditions. Similarly, contemporary announcements of agent a' at date 1 and location 2 have no bearing on the transfer at date 1 and location 1. That is, for this pure exchange environment, an interspatial telecommunication technology would not be used for *contemporary* announcements even though it is allowed. (For more on this, see Section VII below.) On the other hand, the private information, full communication program does allow the announcement of agent a at location 1 and date 1 to enter into the transfer function at location 1 and date 2. Such tie-ins are used because they weaken the damaging effect of incentive compatibility constraints at date 1 for agent a, as can be deduced also by a study of first-order conditions. A similar tie-in is allowed for announcements of agent a' at location 2 across dates 1 and 2.

These tie-ins make the optimal arrangement quadrilateral rather than bilateral. That is, apart from coordination in the choice of allocation rules over the two locations and two dates, as determined by the Pareto weights λ^j, $j=a,a',b$ and b', the solutions to the full information program, solutions to program 3 without the incentive constraints, can be implemented as a sequence of bilateral arrangements, one at each location and date. With tie-ins, however, agent a's announcment at date $t=1$ matters for both agent b at date $t=1$ and for agent b' at date $t=2$. Put crudely, agent a can "borrow" from agent b and promise to "pay" agent b', and similarly for agent a' in his dealings with agents b' and b. In fact, agent b could end up "paying" twice, if he were a "lender" to agent a at date $t=1$ and "pays" back a "loan" from agent a' at date 2. But of course these and all other possibilities are weighted optimally *ex ante* in the determination of the social optimum.

An easy way to deliver explicit examples of beneficial quadrilateral arrangements is to trick the relatively complicated program 3 into looking like program 2. To do this, suppose agents a and a' have identical utility functions, identical endowments, and suffer the same probability distribution determining preference shocks. Similarly, suppose agents b and b' have identical utility functions and identical endowments. Finally, restrict attention to Pareto weights λ^j with $\lambda^a=\lambda^{a'}$ and $\lambda^b=\lambda^{b'}$. Then, as can be deduced formally by an examination of the necessary and sufficient first-order conditions, a solution to program 2 can be viewed as a double solution to program 3, with the probabilities of consumptions to agent a identical with the probabilities of consumptions to agent a', and so on. Thus, Tables 3 and 4 are examples of beneficial quadrilateral arrangements with the understanding, though it does not appear in the notation, that agent a is dealing with agent b at date 1 but dealing with agent b' at date 2, and similarly for agent a' in his dealings with agents b' and b. This will be exploited in what follows.

Perhaps it bears repetition, however, that the perfect costless enforcement premise still

underlies the analysis. That is, all agents are required to show up at the locations specified in Table 5, even though agents b and b' are traveling. Similarly, agents a and b must abide by the prespecified allocation rule at location 1 and at date 1, even though neither will deal with the other ever again. One wonders about collusion or default, even though, as before, one can tell a physical story to support the full commitment program. This subject is of sufficient interest that it is attacked head-on in Section IX below, providing an alternative rationale for communication that is ignored in the intermediate sections which follow.

IV. The Limitations of Oral Assignment Systems in Spatial Settings

The next step in the consideration of communication-accounting systems is to retain the entire setup of Section III, but to limit the communication technology. The first technology to be considered is the most primitive of technologies, namely oral assignment. That is, agents a and a' can make announcements at each date and location of contemporary values for preference shocks and, where relevant, past values as well. But no other record-keeping device is available. That is, agents cannot carry commodities or tokens of any kind, cannot carry written messages, and cannot access some centralized telecommunication record-keeping system. Under these circumstances, only the contemporary state matters for the announcement of agent a or agent a', even though announcements of past histories might be permitted. And thus, by familiar, revelation principle arguments, agents a and a' may without loss of anything essential be restricted to announcing relevant contemporary values (histories can be announced only to the extent that they help to determine contemporary values).

The effect of course is to preclude direct intertemporal tie-ins of the type already analyzed. That is, program 2 reduces to two separate versions of program 1 (with identical Pareto weights in each period), and this is obviously Pareto inferior. Supposing that agents b and b' are constrained to the utility of autarky, for example, agents a and a' suffer.

In fact, though the analysis is somewhat contrived, one can see from this example how spatial organization itself can depend on communication technologies. In particular, suppose that the motive for travel of agent b to agent a' and of agent b' to agent a at the end of date 1 is that the match between a and b and the match between a' and b' deteriorate over time, mimicked by the assumption in the model that K units of the consumption good disappear from the social endowment available at each location at date 2 if agents remain paired. (Alternatively, in a more elaborate model, imagine there are gains to specialization and trade if agents move about.) Then there would be a nontrivial choice among organizations: if agents remain paired, the cost of a deteriorated match must be weighed against the gain from direct intertemporal tie-ins. In fact, agents would choose *ex ante* to remain paired for at least some nontrivial values of cost K. Yet they would not remain paired under perfect costless telecommunications, or even under some of the more restricted systems considered below.

V. Portable Concealable Objects as Record-Keeping Devices

An improved communication technology would allow agents a and a' to carry with them otherwise valueless tokens. In principle, these might be concealed at date 2, but, on the other hand, they might be displayed as a record of things past. In particular, if the allocation of tokens at date 1 is under the complete control of the predetermined allocation rules, then past announcements can be indicated, allowing some of the needed intertemporal tie-ins.

It is possible, in fact, that just one kind of token can allow recovery of the solution to the full communication, private information program, program 2 (actually program 3). In particular, for the environment generating Table 3, date 1 consumptions of agent a are ordered by values of θ_1^a, and the families of date 2 consumptions are ordered in reverse by these values. That is, the date 2, $\theta_1^a = .6$

branch is uniformly lower than the date 2, $\theta_1^a = .5$ branch, which in turn is uniformly lower than the date 2, $\theta_1^a = .4$ branch. Thus, if there were no record of first-period announcements, agent a at date 2 would prefer the $\theta_1^a = .4$ branch no matter what date 2 shock he experiences, and so on. That is, he would claim the highest branch. With tokens, claims can be limited to branches consistent with agent a's display of tokens. And, if agent a is given the least amount of tokens at date 1 for $\theta_1^a = .6$, and the most for $\theta_1^a = .4$, with $\theta_1^a = .5$ in between, then no tokens will be concealed, and the full communication solution will be effected.

Interestingly enough, the full communication solution cannot always be effected with one kind of token, as the next section illustrates.[4]

[4]On the other hand, Douglas Diamond has suggested the following ingenious scheme with the idea that even one token might not be needed for the environment of this section and that oral communication might suffice. Suppose all four agents, a, a', b, and b' are to agree a priori to implement a social optimum as if there were full communication. In particular, agents b and b' are to agree a priori, *in private*, that if agent a reports θ' at date $t = 1$, then agent b is to say some prespecified password, for example, "midnight," whereas if agent a reports θ'', agent b is to say some distinct password, for example, "high noon." Upon b''s arrival to location 1, agent a is to repeat the password, out of countless thousands of possibilities. Agent b' will then know the history of actual past announcements of agent a and is to implement the specified transfer. One problem with this scheme is that the initial agreement between b and b', on passwords as a function of θ announcements, must be *private* between b and b' (if agent a knows the agreement, he will always say the password at date $t = 2$, which allows him to receive the consumption good). Therefore, agents b and b' could just as easily commit themselves to a degenerate password system, always saying the same thing, and planning matters so that a is never to receive the consumption good at date $t = 2$. That is, since the choice of a password function is private, agents b and b' may be supposed to make the best-unconstrained choice, and a requirement that a particular function be chosen has no force if it is not consistent with incentives. Something akin to this is assumed in the "revelation principle" literature: requirements that agents tell the truth in an announcement game have no force unless agents have an incentive to do so. This example illustrates, however, the importance of the assumption in the paper that rules be agreed to publicly and leads to an exploration of enforcement technologies and definitions of reneging. But this is left as a subject for future research.

VI. Multiple-Portable Tokens and Written Messages

The example displayed in Table 4 is one for which one kind of token is not enough. To see this, suppose agent a is given tokens at date 1 for $\theta_1^a = (.4, .6)$ and is given no tokens at date 1 for $\theta_1^a = (.6, .4)$. Then, if at date 1, $\theta_1^a = (.4, .6)$, and at date 2, $(\delta_x, \delta_y) = (.5, 1.5)$, so that $\theta_2^a = (.3, .6)$, agent a would prefer to show no tokens, would claim he was a $\theta_1^a = (.6, .4)$, and also claim $\theta_2^a = (.4, .6)$. On the other hand, if he were given tokens at date 1 under $\theta_1^a = (.6, .4)$ and not otherwise and if θ_1^a were actually $(.6, .4)$, then he would understate tokens at $\theta_2^a = (.6, .3)$, preferring at date 2 the $\theta_1^a = (.4, .6)$ and $\theta_2^a = (.6, .4)$ outcome.

The intuition behind this result, and the contrast with Table 3, are instructive. In Table 3 agent a is either a "borrower" or a "lender" at date $t = 1$, in various degrees, in the sense that the direction of the transaction is reversed at date 2. In Table 4 there are two goods, and agent a can be a "borrower" or a "lender" in either good. Still, "preference reversal" shocks at date 2 can cause agent a to want to pretend to have been a lender in the commodity he did not lend. And this can happen no matter which commodity was lent at date $t = 1$. Of course, two kinds of tokens circumvent this problem, one for each commodity which can be lent. If green tokens are handed out at date 1 and $\theta_1^a = (.4, .6)$ and red at $\theta_1^a = (.6, .4)$, then a display of red tokens could be required at date 2 when $\theta_1^a = (.6, .4)$ is claimed at date 2, and this is not possible if in fact $\theta_1^a = (.4, .6)$ was claimed at date 1.

It might seem from this that the number of kinds of tokens needed to support a full communication optimum is related to the number of commodities or the number of shocks. Actually though, the current environment could be expanded to include any finite number of commodities or shocks, as Table 6 illustrates.[5] Here, if θ_1^a takes on the second value (in an ordered set) at date 1, for example, then agent a receives at date 1 two red tokens and $N-1$ green tokens. At date 2 he cannot overstate past θ_1^a values,

[5]I owe this example to Arthur Kupferman.

TABLE 6—A COMPLETE TWO-TOKEN SYSTEM

Values of θ_1^a	No. of Red Tokens	No. of Green Tokens
1	1	N
2	2	$N-1$
3	3	$N-2$
$\vdots$	$\vdots$	$\vdots$
N	N	1

for example, claiming the third value, for he would be short of red tokens, and he cannot understate past θ_2^a values, for example, claiming the first, for he would be short of green tokens. Thus two kinds of tokens are enough to distinguish past histories. (We shall comment on a related issue, privately observed endowments, in Section X below.)

An interesting general issue concerns when a specified number of kinds of tokens will be enough to constitute a full language, that is, enough to achieve a full communication solution. This issue will not be pursued here, apart from noting any system with a full set of tokens would be equivalent with a system with unrestricted fully displayed written messages. That is, written message systems are the limit of concealable token systems and as such do not require a separate analysis.[6]

VII. Electronic Telecommunications—A Dominant Technology

In the explicit four-agent environment of Section III, there was no gain from electronic telecommunications over space at a point in time. But, modifications of the pure risk-sharing environment can provide a motive for such systems. For example, suppose the consumption good at date 1 can be stored in direct amount K at date 1, carried over without depreciation at the same location to date 2. Suppose also that the preference shocks of agents a and a' at date 1 are driven by some common component, with idiosyncratic noises, and that the common component persists to some extent into preference shocks at date 2. Then shock $\theta_1^{a'}$ of agent a' can be used to help forecast the marginal utility of agent a at date 2, and a high marginal utility for consumption at date 2 would motivate relatively high investment K at location one, date 1. A similar argument applies for shock θ_1^a at location two and date one.

To check on this logic, it is useful to go through the formal exercise of writing down a programming problem for the determination of Pareto-optimal allocations with electronic, interspatial telecommunications and storage. In particular, invoking the kind of symmetry assumptions used earlier to trick the four-agent program to a two-agent program, let $\pi_1(c, K|\theta_1^a, \theta_1^{a'})$ denote the probability of consumption bundle c *and* storage K given announced (and actual) $\theta_1^a, \theta_1^{a'}$ values, where, given K, bundle c lies in the set $\{(c^a, c^b)\colon\ c^a + c^b \le e^a + e^b - K\}$. Similarly, let $\pi_2(c, K|\theta_1^a, \theta_1^{a'}, \theta_2^a)$ denote the probability of consumption bundle c *and* storage K given the specified triple of announced (and actual) parameter values, where, for given K, the bundle c lies in $\{(c^a, c^b)\colon\ c^a + c^b \le e^a + e^b + K\}$. Then the program for the determination of a full communication private information optimal arrangement is

Program 4: Maximize by choice of the $\pi_1(\cdot)$ and $\pi_2(\cdot)$ the objective function

$$
\begin{aligned}
(9)\quad & \lambda^a\Big\{\sum_{\theta_1^a}\sum_{\theta_1^{a'}} p(\theta_1^a, \theta_1^{a'}) \\
&\quad \times \sum_K \sum_c \pi_1(c, K|\theta_1^a, \theta_1^{a'}) U^a(c^a, \theta_1^a) \\
&\quad + \beta \sum_{\theta_1^a}\sum_{\theta_1^{a'}}\sum_{\theta_2^a} p(\theta_2^a|\theta_1^a, \theta_1^{a'}) p(\theta_1^a, \theta_1^{a'}) \\
&\quad \times \sum_K \sum_c \pi_2(c, K|\theta_1^a, \theta_1^{a'}, \theta_2^a) U^a(c^a, \theta_2^a)\Big\} \\
& \lambda^b\Big\{\sum_{\theta_1^a}\sum_{\theta_1^{a'}} p(\theta_1^a, \theta_1^{a'}) \sum_K \sum_c \pi_1(c, K|\theta_1^a, \theta_1^{a'}) U^b(c^b) \\
&\quad + \beta \sum_{\theta_1^a}\sum_{\theta_1^{a'}}\sum_{\theta_2^a} p(\theta_2^a|\theta_1^a, \theta_1^{a'}) p(\theta_1^a, \theta_1^{a'}) \\
&\quad \times \sum_K \sum_c \pi_2(c, K|\theta_1^a, \theta_1^{a'}, \theta_2^a) U^b(c^b)\Big\},
\end{aligned}
$$

[6]More generally, a token system is a kind of written message system, a limited one if combinations of tokens cannot adequately convey past history. The idea that tokens are related to writing as "words" are to language may be familiar—real languages actually evolved from token accounting systems. See Schmandt-Besserat (1979).

subject to consistency in the choice of K, namely, for all $\theta_1^a, \theta_1^{a'}, \theta_2^a$

$$(10)\qquad \sum_c \pi_2(c, K|\theta_1^a, \theta_1^{a'}, \theta_2^a) \equiv \sum_c \pi_1(c, K|\theta_1^a, \theta_1^{a'}),$$

subject to incentive constraints for agent a at date $t=2$, for all past announcements $\theta_1^a, \theta_1^{a'}$, and for all actual θ_2^a and counterfactual $\tilde{\theta}_2^a$,

$$(11)\qquad \sum_c \sum_K U^a(c^a, \theta_2^a)\pi_2(c, K|\theta_1^a, \theta_1^{a'}, \theta_2^a) \geq \sum_c \sum_K U^a(c^a, \theta_2^a)\pi_2(c, K|\theta_1^a, \theta_1^{a'}, \tilde{\theta}_2^a),$$

subject to incentive constraints for agent a at date $t=1$, for all actual θ_1^a and counterfactual $\tilde{\theta}_1^a$

(12)

$$\sum_{\theta_1^{a'}} p(\theta_1^{a'}|\theta_1^a)\sum_c \sum_K U^a(c^a, \theta_1^a)\pi_1(c, K|\theta_1^a, \theta_1^{a'}) + \beta\sum_{\theta_1^{a'}}\sum_{\theta_2^a} p(\theta_2^a, \theta_1^{a'}|\theta_1^a) \times \sum_c \sum_K U^a(c^a, \theta_2^a)\pi_2(c, K|\theta_1^a, \theta_1^{a'}, \theta_2^a)$$

$$\geq \sum_{\theta_1^{a'}} p(\theta_1^{a'}|\theta_1^a)\sum_c \sum_K U^a(c^a, \theta_1^a)\pi_1(c, K|\tilde{\theta}_1^a, \theta_1^{a'}) + \beta\sum_{\theta_1^{a'}}\sum_{\theta_2^a} p(\theta_2^a, \theta_1^{a'}|\theta_1^a) \times \sum_c \sum_K U^a(c^a, \theta_2^a)\pi_2(c, K|\tilde{\theta}_1^a, \theta_1^{a'}, \theta_2^a).$$

The incentive constraints are much as before except that agent a takes as given that agent a' is announcing truthfully at date 1, and agent a learns this parameter announcement *after* the allocation is effected at date 1. Constraint (10) ensures that the choice of K is the same whether viewed as determined by the allocation rule $\pi_1(\cdot)$ or the allocation rule $\pi_2(\cdot)$. Alternatively, and more naturally, one could have let the allocation rule $\pi_2(c|\theta_1^a, \theta_1^{a'}, \theta_2^a, K)$ be conditioned on K, but then it would not have been obvious that the essential program is linear.

VIII. Portable Messages as a Location Assignment Device

We have now passed through the gamut of communication technologies in the context of the same four-agent, two-location model. Each of the communication technologies considered in the model played a role as a device for keeping track, if possible, of past announcements by the agents with privately observed preference shocks, agents a and a'. No records were needed of the histories experienced by agents b and b'. In contrast, this section shows how records may be needed for travelers who experience no direct shocks *if* their assignments to locations is endogenous, part of the (optimal) social mechanism.

The model is modified by supposing there is only one traveler, agent b, who is paired initially at date $t=1$ with an agent a, experiencing preference shock θ_1^a, then is paired at date $t=2$ either with an agent d, who is to experience shocks θ_2^d, or with an agent e, who is to experience shocks θ_2^e, but is never paired to both. Further, shocks θ_1^a contain information on forthcoming θ_2^d and θ_2^e, so that an assignment at date $t=1$ matters. That is, assignment at date $t=1$ is a nontrivial function of announced (and actual) θ_1^a. For simplicity, suppose agent a cares about date $t=1$ consumption only and that agents d and e care about date $t=2$ consumption only, though agent b, the traveler, cares about consumption at both dates. Then, to ensure some mutual beneficial trade, suppose there are two goods at each date. The notation for endowments e_t^i is as before.

Letting variable l denote the location assignment of agent b at the end of date 1, to either agent d or e, that is either $l=d$ or $l=e$, the programming problem for the determination of a *full communication* private information optimum is

Program 5: Maximize by choice of date $t=1$ (potentially random) consumption and as-

signment rules, $\pi_1(c|\theta_1^a)$ and $\pi_1(l|\theta_1^a)$, respectively, and by choice of date $t=2$ consumption rules, $\pi_2(c|\theta_2^d, l=d)$ and $\pi_2(c|\theta_2^e, l=e)$, the objective function

$$(13)\quad \lambda^a\Big\{\sum_{\theta_1^a} p(\theta_1^a)\sum_c \pi_1(c|\theta_1^a)U^a(c,\theta_1^a)\Big\}$$

$$+\lambda^b\Big\{\sum_{\theta_1^a} p(\theta_1^a)\sum_c \pi_1(c|\theta_1^a)U^b(c)+\beta\sum_{\theta_1^a} p(\theta_1^a)$$

$$\times\Big[\pi_1(l=d|\theta_1^a)\sum_{\theta_2^d} p(\theta_2^d|\theta_a^a)\sum_c U^b(c)\pi_2(c|\theta_2^d,l=d)$$

$$+\pi_1(l=e|\theta_1^a)\Big\{\sum_{\theta_2^e} p(\theta_2^e|\theta_1^a)\sum_c U^b(c)\pi_2(c|\theta_2^e,l=e)\Big]\Big\}$$

$$+\lambda^d\Big\{\sum_{\theta_1^a} p(\theta_1^a)\Big[\pi_1(l=d|\theta_1^a)\sum_{\theta_2^d} p(\theta_2^d|\theta_1^a)$$

$$\times\sum_c U^d(c,\theta_2^d)\pi_2(c|\theta_2^d,l=d)$$

$$+\pi_1(l=e|\theta_1^a)\sum_{\theta_2^d} p(\theta_2^d|\theta_1^a)U^d(e_2^d,\theta_2^d)\Big]\Big\}$$

$$+\lambda^e\Big\{\sum_{\theta_1^a} p(\theta_1^a)\Big[\pi_1(l=d|\theta_1^a)\sum_{\theta_2^e} p(\theta_2^e|\theta_1^a)U^e(e_2^e,\theta_2^e)$$

$$+\pi_1(l=e|\theta_1^a)\sum_{\theta_2^e} p(\theta_2^e|\theta_1^a)\sum_c U^e(c,\theta_2^e)\pi_2(c|\theta_2^e,l=e)\Big]\Big\},$$

subject to incentive constraints for agent a at date $t=1$, and incentive constraints for agents e and d at date $t=2$. Formally, program 5 can be converted to a linear program, for example by letting

$$(14)\quad \pi_2(c,l=d|\theta_1^a,\theta_2^d)$$
$$\equiv \pi_2(c|\theta_2^d,l=d)\pi_1(l=d|\theta_1^a),$$

and similarly for $\pi_2(c,l=e|\theta_1^a,\theta_2^e)$; by imposing a linear equation to ensure that the left-hand side of (14) is independent of θ_2^d, for example; and by imposing consistency with $\pi_1(l=d|\theta_1^a)$, as in equation (10), program 4.

With limited communication, an otherwise optimal assignment of agent b to either agent d or to agent e may not be assured. Announced shock θ_1^a at date $t=1$, and, in shorthand notation, assignment $l(\theta_1^a)$ at date $t=1$, are private to agent b when b meets agent d or agent e at date $t=2$.

TABLE 7—OPTIMAL LOCATION ASSIGNMENT

(θ_x^a,θ_y^a)	(c_x^a,c_y^a)	$l(\theta_1^a)$	(θ_x^d,θ_y^d)	(c_x^d,c_y^d)	(θ_x^e,θ_y^e)	(c_x^e,c_y^e)
(.4,.6)	(2,6.31)	d	(.7,.5)	(10,3)		
			(.3,.5)	(2,6.24)		
(.6,.4)	(6,2)	e			(.7,.5)	(10,3)
					(.3,.5)	(2,6.24)

An example helps to make the point. In particular, suppose agent b has preferences of the form $U^b(c_x,c_y)=c_x+c_y$ at each date, with discount rate $\beta=.9$. Preferences of agent a at date 1 are of the form

$$U^a(c_x,c_y,\theta_1^a)=(c_x)^{\theta_x^a}+(c_y)^{\theta_y^a},$$

with $\theta_1^a=(\theta_x^a,\theta_y^a)$ either (.4,.6) or (.6,.4), each drawn with probability 1/2. Preferences of agents d and e at date 2 are of a similar form to agent a's with correlated parameter draws, that is,

$$\theta_1^a=(.4,.6)$$
$$\Rightarrow\begin{cases}\theta_2^d=(.7,.5) \text{ and } \theta_2^e=(.3,.5) & \text{with Prob .8}\\ \theta_2^d=(.3,.5) \text{ and } \theta_2^e=(.7,.5) & \text{with Prob .2}\end{cases}$$

$$\theta_1^a=(.6,.4)$$
$$\Rightarrow\begin{cases}\theta_2^d=(.7,.5) \text{ and } \theta_2^e=(.3,.5) & \text{with Prob .2}\\ \theta_2^d=(.3,.5) \text{ and } \theta_2^e=(.7,.5) & \text{with Prob .8}\end{cases}$$

Then the consumption allocations for agents a, d, and e and assignment rule $l(\theta)$ are displayed[7] in Table 7.

When $\theta_1^a=(.4,.6)$, agent a is supposed to be assigned to agent d, and when $\theta_1^a=(.6,.4)$, agent a is supposed to be assigned to agent e, each assignment consistent with the likelihood of high marginal utility for agent d or agent e, respectively. But agent b, anticipating high transfers under the full communication-assignment rule, wants to go in just the contrary direction.

[7]Endowments are five uniformly, and $\lambda^a=\lambda^d=\lambda^e$ with the utility of agent b at autarky.

However, if agent b carries a token of one kind when $l(\theta_1^a) = d$ and a token of another kind when $l(\theta_1^a) = e$; if he must be with agent a at date $t = 1$, and then with agent d or agent e at date $t = 2$; and if he must abide by the prespecified allocation rule when paired, then a failure to display tokens of the right kind can indicate that agent b is not abiding by the rules. Curiously, two kinds of tokens are needed. For suppose there were only one kind of token in use and agent b were given it for $\theta_1^a = (.4,.6)$ and not when $\theta_1^a = (.6,.4)$, for example. Then when $\theta_1^a = (.6,.4)$, agent b wants to go to agent d where the likelihood of high transfers is less. But he has no tokens to show, so he must go to agent e. But when $\theta_1^a = (.4,.6)$, agent b wants to go to agent e and can do so by concealing his tokens. Again, two kinds of tokens circumvent this problem.

IX. Portable Messages as a Device for Enforcement

The structure of Section VIII takes as given that agent b must show up and be paired with agent a at date $t = 1$ and must choose to be paired with either agent d or agent e at date $t = 2$. A slight weakening of the commitment premise would allow agent b to choose whether to participate at some dates. In fact, spatial separation alone could then imply private information in the sense that failure of one agent to participate with a second agent at a given location in a given date might not be known by a third party at a subsequent date. That is, communication can be valuable if commitment is limited, even with no underlying uncertainty.

The delicate part of schemes with limited commitment is that one can undo the structure of the problem altogether, so that *ex ante* agreements have no force at all. In that case, there would be nothing to communicate since plans have no content, are known to have no content, and so on. Thus, *some* prior commitment is needed. In this section one starts with essentially full commitment, supposing that if agents show up at a prespecified location they *must* abide by the prespecified allocation mechanism in place at that location. Further, at the very last date, agents *must* show up at prespecified locations, no matter what. Hence, the (only) aspect of limited commitment which is introduced is the idea that each agent can choose at all dates but the last whether to participate under prespecified allocation rules. The alternative to participation is to eat one's endowment.

For clarity, we return to the basic four-agent, two-location, two-period model of the paper, but eliminate shocks to preferences. Otherwise, the structure is as before. And as before it is best to begin with the full communication record-keeping technology, supposing actions at locations 1 and 2 at the first date are fully recorded and can be used in the allocation rules at date $t = 2$. The objective in the planning period is to choose (deterministic) allocations c_t^i, $i = a, a', b, b'$, $t = 1,2$, to maximize a weighted sum of discounted utility streams across the four agents, namely

$$(15) \qquad \sum_i \lambda^i \Big\{ \sum_{t=1}^2 \beta^{t-1} U^i(c_t^i) \Big\},$$

subject to the obvious resource constraints

$$(16) \qquad c_t^i + c_t^j = e_t^i + e_t^j,$$

and so on for the relevant agent pairing (i, j) at each date t.

The relevant decision for a given agent at the first date is whether to participate as planned at date $t = 1$; participation cannot be assumed and must be induced. But since constraints which ensure participation at date 1 generally damage the program, relative to the full commitment program, one wants such participation constraints to be as weak as possible. This is done by making the "penalty" at date 2 for failure to participate at date 1 as strong as possible. That is, the first-period participation constraint for agent i takes the form

$$(17) \qquad U^i(c_1^i) + \beta U^i(c_2^i) \geq U^i(e_1^i) + \beta U^i(0),$$

where $c_2^i = 0$ on the right-hand side of (17) is the obvious penalty, the lower bound on i's consumption at date 2. Again, the interpretation is that if agent i participates at date 1,

receiving c_1^i as planned, then when i shows up at date 2, and he must by assumption do so, he will receive c_2^i as planned, whereas if agent i does not participate at date $t=1$, eating his endowment e_1^i, something which cannot be directly thwarted, then when agent i shows up at date $t=2$, the fact of default is known and he receives zero. In short, the program for the determination of Pareto-optimal allocations is one of maximizing objective function (15) subject to four resource constraints (16) and subject to four participation constraints (17).[8]

To implement a solution to this program *without* costless telecommunicated record keeping, but with portable concealable objects such as tokens, suppose that if agent i participates at date $t=1$ he receives c_1^i *and* a prespecified number of tokens (one will do). Tokens, by assumption, cannot be acquired elsewhere. Then a display of tokens at date 2 effects c_t^i to agent i, and a failure to display tokens effects the penalty, zero consumption. Thus tokens communicate whether or not agent i reneged at date 1, an action which is otherwise unknown to agent i's trading partner at date 2.

This argument can be extended to any T-period, finite horizon model with choice of participation at all dates t but the last. There would be a participation constraint for each agent i at each date t, $t=1,2,\dots,T-1$, and the constraint for agent i at date t would be[9]

$$(18)\quad \sum_{s=t}^{T} \beta^{(s-1)}U^i(c_s^i) \geq \sum_{s=t}^{T-1} \beta^{(s-1)}U^i(e_s^i)+\beta^{T-1}U^i(0).$$

With discount rate β less than one, participation at relative early stages is achieved not by high penalties at the last date T but by the gain from the enduring relationship, that is, from participation at intermediate dates. The number of tokens, or length of written messages, needed to implement the limited commitment solution with otherwise limited communication may get large as T goes to infinity.

X. Privately Observed Endowments

The analysis above is now easily generalized to incorporate the case of random and privately observed endowments. Indeed, as with tokens, claimed values of endowments can trigger required displays, some of which would prove infeasible for certain actual realizations of endowments.[10] In short, incentive constraints need be imposed only for claimed endowment vectors which are less than or equal to realized endowment vectors, component by component.

More formally, for the two-agent, one-consumption period model of Section II, let θ^a denote the vector of endowments of agent a, as well as a vector of shocks to the preferences of agent a as before. Also let $\tau=(\tau^a,\tau^b)$ denote a vector of transfers from agents a and b, supposing for simplicity a finite number of possible values for these transfers. Then let $\pi(\tau|\theta^a)$ denote the probability of transfer τ conditioned on announcement θ^a, so that for each θ^a, $\pi(\tau|\theta^a)$ is a lottery over τ values satisfying $\theta^a-\tau^a \geq 0, e^b-\tau^b\geq 0, \tau^a+\tau^b=0$. The program for the determination of Pareto optimal allocations is then

Program 6: Maximize by choice of the $\pi(\tau|\theta^a)$ the objective function

$$(19)\quad \lambda^a\Big\{\sum_{\theta^a} p(\theta^a)\sum_{\tau} U^a[\theta^a-\tau^a,\theta^a]\pi(\tau|\theta^a)\Big\} + \lambda^b\Big\{\sum_{\theta^a} p(\theta^a)\sum_{\tau} U^b[e^b-\tau^b]\pi(\tau|\theta^a)\Big\},$$

[8]More generally, we can write down a program which lets participation or assignment at date 1 be a choice variable, subject to constraint that if the assignment is not to participate, then agent i receives his endowment, and also subject to obedience constraints, that if agent i is assigned to participate, then he must prefer to do so, and if he is assigned not to participate, then he must prefer not to do so. It can be established that for any solution with nonparticipation as the assignment, there is a utility equivalent solution with participation as the assignment, hence the program given above.

[9]That this is without loss of generality can be established as in fn. 8.

[10]It might be noted that much of the literature on resource allocation mechanisms ignores privately observed endowments. Important exceptions are Andrew Postlewaite (1974), Hurwicz, Eric Maskin, and Postlewaite (1980), and also Pipat Pithyachariyakul (1982).

subject to incentive constraints, for every $\tilde{\theta}^a \leq \theta^a$,

$$(20)\quad \sum_{\tau} U^a[\theta^a - \tau^a, \theta^a]\pi(\tau|\theta^a) \geq \sum_{\tau} U^a[\theta^a - \tau^a, \theta^a]\pi(\tau|\tilde{\theta}^a).$$

Again, solutions to program 6 can require actual displays. For example, with one good, if a high endowment θ^a for agent a is associated with a high marginal utility shock θ^a for agent a, then the *ex ante* optimal insurance solution may have agent a receiving the good when his endowment is high, after it is displayed, and surrendering it otherwise, when his endowment is low.[11] Without pretransfer displays this would not be incentive compatible, for agent a would always claim to have a high endowment, in order to receive the higher net transfer.

Generally, though, pretransfer displays cannot completely overcome the incentive problems of private information, and the analysis of communication-accounting systems with privately observed endowment shocks would proceed along the lines of the paper as for the case of privately observed preference shocks. But a new and interesting case for analysis now emerges, the case of commodity storage coupled with the presence of *no* intrinsically useless (storable) tokens. In this case, displays of otherwise private stocks can serve in part as worthless tokens did to reveal past histories. But the communication-accounting system with these bonafide commodity tokens would be more limited than the system with intrinsically useless tokens.

Consider the four-agent, two-location, two-period economy, tricked into the two-agent, one-location, two-period economy by the symmetry conditions, and suppose no communication system is available other than that achieved by direct commodity storage, publicly observed at the time storage decisions are taken but not necessarily observed later. Then let $\pi(\tau, K|\theta_1^a)$ denote a lottery over transfers $\tau = (\tau^a, \tau^b)$ and storage in amount K at date 1, conditioned on announcement θ_1^a by agent a at date 1, with transfers τ and storage K satisfying $\theta_1^a - \tau^a \geq 0$, $e^b - \tau^b \geq 0$, $K = \tau^a + \tau^b$. Also, let $\pi_2(\tau|\theta_2^a + K)$ denote a lottery over transfers $\tau = (\tau^a, \tau^b)$ at date 2, conditioned on announcement $\theta_2^a + K$ by agent a at date 2, with $\theta_2^a + K - \tau^a \geq 0$, $e^b - \tau^b \geq 0$, $\tau^a + \tau^b = 0$.

The point is that the natural state variable at date 2 is $\theta_2^a + K$; only this *sum* can enter into the allocation rule at date 2. Thus if endowment θ_2^a were constant, storage K, in varying with endowment θ_1^a, might allow a complete revelation of past histories, as for the earlier examples with intrinsically useless tokens. But with endowment θ_2^a nonconstant, a display of goods may be possible, even when commodity storage itself is inadequate. That is, in an effort to achieve the former allocation, agent a may occasionally have the ability to display when formerly he did not, and this would undercut the information revelation role of tokens. On the other hand, if storage K were constant, independent of endowment θ_1^a, then allocations can be indexed by θ_2^a. But with K nonconstant, depending on θ_1^a, the θ_2^a-information role of contemporary displays is mitigated. In general, then, neither complete past histories nor contemporary states will enter into the de facto allocation rules at date 2. And further constraining the information role of commodity tokens is the fact that commodity storage K as an information variable distorts intertemporal allocations from what they would have been with costless, intrinsically useless tokens.

[11] This example may seem somewhat contrived, but it can be given a more compelling interpretation. Briefly, imagine that agents a and b each have an investment project and that returns from investment projects at date 1 must be reinvested at date 1 in order to yield an idiosyncratic, household-specific, nontraded consumption good at date 2. The investment good can be reallocated at date 1, however, and would be reallocated in an *ex ante* optimal arrangement if high investment good returns at date 1 were indicative of high consumption returns at date 2, yielding high marginal return of the *indirect* utility function for contemporary investment. For further details, see the working paper, Townsend (1986).

REFERENCES

Arrow, Kenneth J., "Le role des valeures boursierés pour la repartition la meilleure des risques," *Econometrie*, 1953, 41–48.

Baric, Lorraine, "Some Aspects of Credit, Saving, and Investment in a Nonmonetary Economy," (Rossel Island), in R. Firth and B. Yamey, eds., *Capital, Saving, and Credit in Peasant Societies*, Chicago: Aldine, 1964.

Brunner, Karl and Meltzer, Allan H., "The Uses of Money: Money in the Theory of an Exchange Economy," *American Economic Review*, December 1971, *61*, 784–805.

Debreu, Gerard, *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, New York: Wiley & Sons (for Cowles Foundation), 1959.

Firth, Raymond, *Primitive Polynesian Economy*, London: G. Routledge & Sons, 1939.

Gale, Douglas, "Money, Information and Equilibrium in Large Economies," *Journal of Economic Theory*, August 1980, *23*, 28–65.

Harris, Milton and Townsend, Robert M., "Resource Allocation under Asymmetric Information," *Econometrica*, January 1981, *49*, 33–64.

Hurwicz, Leonid, "On Informationally Decentralized Systems," *Decision and Organization*, in C. G. McGuire and R. Radner, eds., Amsterdam: North-Holland, 1972, Ch. 14.

______, **Maskin, Eric and Postlewaite, Andrew**, "Feasible Implementation of Social Choice Correspondences by Nash Equilibria," unpublished manuscript, University of Minnesota, August 1980.

Malinowski, B., *Argonauts of the Western Pacific*, New York: Dutton, 1953.

Mount, Kenneth and Reiter, Stanley, "The Informational Size of Message Spaces," *Journal of Economic Theory*, June 1974, *8*, 161–92.

Myerson, Roger, "Incentive Compatibility and the Bargaining Problem," *Econometrica*, January 1979, *47*, pp. 61–73.

Ostroy, Joseph M., "The Informational Efficiency of Monetary Exchange," *American Economic Review*, September 1973, *63*, 597–610.

______ **and Starr, Ross M.**, "Money and the Decentralization of Exchange," *Econometrica*, November 1974, *42*, 1093–113.

Pithyachariyakul, Pipat, "Exchange Mechanism, Strategies, and Efficiency with Private Information," unpublished doctoral dissertation, Northwestern University, August 1982.

Schmandt-Besserat, Denise, "The Earliest Precursor of Writing," *Scientific American*, September 1979, *241*.

Townsend, Robert M., "Optimal Multiperiod Contracts and the Gain from Enduring Relationships under Private Information," *Journal of Political Economy*, December 1982, *90*, 1166–86.

______, "Economic Organization with Limited Communication," Working Paper No. 86-3, April 1986.

Usher, Abbott P., *The Early History of Deposit Banking in Mediterranean Europe*, Cambridge: Harvard University Press, 1943.

Women's Work, Sibling Competition, and Children's School Performance

By FRANK P. STAFFORD*

This research offers an interpretation of the relationship between fertility, child spacing, family resources, and market work by mothers, and the subsequent cognitive skills of grade schoolers as reported by their teachers. First, an approach to fertility as the outcome of a deliberate choice process is developed, and then empirical evidence from a small panel of U.S. households first interviewed in 1975–76 when they had preschool children is presented. In a 1981–82 reinterview, a supplementary project was that of obtaining teacher ratings of school performance of individual children.

Large family size, as measured by the number of siblings of given age and sex in the household during preschool years (in 1975–76), has an important, negative impact on the child's subsequent grade-school performance. Boy siblings in nearby age ranges have the most negative impact on performance while teenage siblings of either sex have no systematic adverse effect. Parental resources, as measured by income, education, child-care time, and a mother's reduced market time are associated with greater cognitive skills, and can offset the apparent disadvantage of having siblings in nearby age intervals.

There appears to be a significant tradeoff between a market career and a home career for women. Women who have more children spaced over wider age intervals and who devote more time to child care and less to market work presumably get more benefits from their home career in the form of enhanced child development. On the other hand, full-time market work (i.e., hours in the labor market for pay) is important for earnings growth of the mother (Mary Corcoran, Greg Duncan, and Michael Ponza, 1983), and family income has a favorable effect on school performance, so the apparent choices facing women are more equivocal.

I. A Conceptual Framework

The rate of per capita development of the children, $\dot{K}$, is given by

$$(1) \qquad \dot{K} = Q(c, t; n, a)$$

$$Q_c > 0, \quad Q_{cc} < 0, \quad Q_t < 0$$

$$Q_n < 0$$

$$n \geq 1,$$

where $Q(\cdot)$ is a function that describes the relationship between child development and parental time inputs, c; age of child or time, t; number of children in the household, n; and ability or developmental potential of the children, $\underline{a}$. In this formulation, large values of $\underline{a}$ lower the cost of adding to child development.

The development of market skills of the parent, $\dot{S}$, is given by

$$(2) \qquad \dot{S} = f(S, m) - \delta S \; f_m, f_S > 0;$$

$$f_{mm}, f_{SS} < 0,$$

where $f(\cdot)$ is a function relating one's existing stock of skills, S, and one's training time, m, to the growth of market skills, and

*University of Michigan, Ann Arbor, MI 48109.
Thanks go to journal referees and seminar participants at Cornell and the Universities of Chicago, Michigan, and Saarlandes (West Germany) for comments on an earlier draft. I would also like to thank the Universitat des Saarlandes for support in completing the work on this project. Special thanks go to Ted Bergstrom, Charles Brown, John Bound, Greg Duncan, Jacque Eccles, Daniel Hill, Tom Juster, and David Lam for particularly helpful suggestions. Linda Young and Dorothy Kempter provided essential advice on data and computational issues.

δ is a depreciation rate. Total time available $(=1)$ can be divided between child care, c; training, m; and market work, l;

$$(3) \qquad 1 \geq l + m + c.$$

Welfare of the parent, V, is a function of earnings over the entire planning period, and the children's development state at the end of the planning period.

$$(4) \qquad V = \int_0^T e^{-pt} \alpha l S dt + K_T,$$

where T is the end of the planning period and p is the discount rate. The rental rate or per unit wage rate on market skills is α, so that αls is the income flow from labor market activity. The initial level of earnings capacity at the beginning time (0) is denoted as $S(0) = S_0$.

Because children's well-being is built up via (1), at the end of the planning period it is given by

$$(5) \qquad K_T = \int_0^T Q(c, t; n, a)\, dt.$$

Substitution of (5) into (4) and maximization of (4) subject to (1), (2) and (3) leads to first-order conditions

$$(6) \qquad Q_c(t) = e^{-pt} \alpha S = \lambda_S f_m,$$

where λ_S is the shadow value of (additional) market skills at time t and the time derivative is given as

$$(7) \qquad \dot{\lambda}_S = -e^{-pt} \alpha l - \lambda_S f_S + \delta.$$

Full solution of the model can be developed only for more restricted functional forms of $Q(\cdot)$, $V(\cdot)$, and $f(\cdot)$. Although such analysis cannot be unthinkingly generalized, it is highly probable that some results are robust in the face of alternative functional forms. Specifically, it is not difficult to imagine that the path of λ_S satisfying conditions for a maximum in (7) will decline through time, possibly after an initial rise. Further, by assumption of time-varying elements in the production of child care in which early inputs have larger developmental payoffs, the effective policy will involve substantial reductions in market work early in the planning period when young children are present. In this case the early life cycle will be devoted primarily to child care and training. The marginal value of time devoted to child care is equated to the marginal payoff to market time, which is in turn equated to the marginal payoff to training time as can be seen in (6).

A main point of the model set out above is to highlight lifetime choices of market career versus child care. To illustrate this more explicitly, we can simplify (1) as follows to represent total child-care time rather than both number of children and care per child

$$(8) \qquad \dot{K} = \gamma \ln c, \quad o < \gamma < 1.$$

The growth of market skills is given as

$$(9) \qquad \dot{S} = A(mS)^{\beta} - \delta S, \qquad o < \beta < 1 \quad \text{and} \quad o < \delta < 1.$$

The role of greater market potential can be examined by considering larger values of A or β (assuming $mS > 1$) while greater child-care potential can be examined by considering larger values of γ.

The interior solution to this problem can be represented by

$$(10) \quad \dot{\lambda}_S = \lambda_S \delta - \lambda_R \alpha + \gamma / S$$

$$(11) \quad \dot{S} = A[\beta \lambda_S A / \lambda_R \alpha]^{\beta/(1-\beta)} - \delta S,$$

where λ_R is the shadow value of market earnings in the lifetime plan (Harl Ryder, Frank Stafford, and Paula Stephan, 1976). Figure 1 portrays the interior solution for two people with differing child-care potential $(\gamma_1 > \gamma_2)$ and differing market potential $(A_1 < A_2)$.

The path labeled 1 has a less extensive buildup of market capital and is closer to the unit child-care locus $(\hat{c} = 1)$ throughout,[1] im-

[1]The $\hat{c} = 1$ locus is derived by noting that the necessary conditions for an optimum imply that $c = \hat{c} = \gamma / \alpha_S$. Therefore, $\hat{c}$ is a decreasing function of S and is independent of λS.

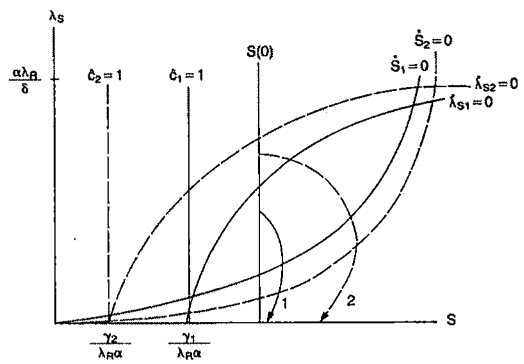


FIGURE 1. LIFE-CYCLE CHILD CARE AND MARKET CAREER

plying a much more lifelong involvement with child care. In contrast, path 2 has a more extensive buildup of market capital (with greater associated training and market work), moves through the phase plane more quickly to higher-market skill levels, and is farther from the $\hat{c}=1$ locus throughout, implying a smaller commitment to child care *and* one which is proportionately concentrated in the early years.[2]

An increase in financial resources can be argued to imply a lower value of additional market income, λ_R, because of diminishing marginal utility of money. This would require a minor modification of (4), which does not alter other aspects of the first-order conditions. In this case, the $\dot{\lambda}_S = 0$ locus is rotated clockwise, the $\dot{S}=0$ locus rises less steeply from the origin, and the speeds of $\dot{\lambda}_S$ and $\dot{S}$ are both increased. The overall qualitative impact on child-care and career paths is difficult to assess. In particular, it does not seem easy to demonstrate that lifetime child-care time will necessarily increase as might be argued in a simple static model in which child well-being is a normal good.

There are several more specialized behavioral questions which we will want to address in the empirical work. Do those with home careers (path 1) choose greater numbers of children at the expense of time resources per child? We also want to examine the impact of market work on child development. Although career people will presumably have fewer resources devoted to child care, will they reduce number of children more than care per child? Does care per child have a major impact on child development?

One way of thinking about the parents' choice over the number of children versus care per child is in the context of the "Rotten Kid Theorem" or RKT (Gary Becker, 1981). Here the family head who regards normal good arguments in his welfare function the resources available to each other family member will, via transfers, create incentives for each of them to act in the common interest; pursuit of own gain which does not increase the total family resources to their potential simply reduces the rotten kid's individual resources.

As has been argued (Theodore Bergstrom, 1987), the RKT holds only if certain conditions are met, including the requirement that the family-utility possibility frontier is a simplex that is subject to parallel shifts by the actions of the "kids." For the case at hand, consider the "parental prodigal parable" (PPP) in which the behavior of the parent can be thought to create a nonhomothetic shift outward in the frontier. In the PPP case the self-indulgent parent would have incentives to pursue behaviors such as market work (and own consumption of goods) or perhaps to have additional children at the expense of the existing family members even if someone else were "in control."

The literature on sociobiology (Richard Dawkins, 1976) also implies a conflict between parents and their offspring. In the "selfish-gene" model parents have an interest in limiting the number of additional children only insofar as the additional children reduce the total number of genes embodied in surviving offspring. The existing children have what is called a coefficient of relatedness (percent of shared genes) to themselves of 1.0 and to their siblings a coefficient of .5. From the parents' perspective, each child has a coefficient of related-

[2]Note that the minor anomaly of increasing child care as T approaches could be corrected by a time varying γ, but this would add unneeded complexity to the problem.

ness of .5. This disparity between the relative weight which parents give to (additional) siblings and that given by the siblings to themselves and additional siblings creates a conflict of interest between the parents and their children. In summary, there are theoretical reasons from both economics and sociobiology to expect a conflict between the interests of the children and their prodigal parents: Children will tend to be made worse off by added siblings.

II. Empirical Analysis

The data are from a national survey of time use conducted in 1975–76 by the Survey Research Center of The University of Michigan (Thomas Juster and Frank Stafford, 1985). A key survey feature was the collection of four time diaries over a period of a year from each household for both head and spouse. The sample for analysis in this paper is small ($N = 77$) because it is limited to those two-parent-present families with preschool children as of 1975–76 who were successfully reinterviewed in 1981–82. Despite the small sample size, the main effects of siblings and maternal time appear to be established, and the cost per data point is sufficiently high (well over $1000 per observation) that alternative data sets are unlikely to be available to test the importance of preschool time allocation and fertility history of the parents.

If families seek to maximize the sum of their children's success without regard to distribution across children, and if success is an increasing function of parental time, they will devote more time to the more competent child, even though marginal input payoffs across children will be equalized. On the other hand, if they seek to achieve equality of child well-being in the face of differences in developmental potential or seek to achieve a performance standard of their children at some norm in comparison to other children in their social reference group, then extraordinary amounts of child care and child-related housework may be an indication that the parents are compensating for perceived developmental weaknesses of one (or more) of their children (Becker and Nigel Tomes, 1979).

From our research perspective, the empirically simplifying assumption is clearly that parents make preschool child input decisions as a function of variables which are reasonably independent of the child's inherent characteristics or eventual rate of development. One could argue that this is because, aside from a few preschool children who seem to have extreme and obvious high- and low-development rates, the parents simply do not know enough to act on their possible systematic preferences for success maximization or equalization of well-being across children. In this setting deviations of time beyond normal levels represent a "treatment" that is both independent of the child's development potential and can be apportioned across two or more children in the multiple-child households. If this is not so, the assignment of parental inputs to each child in a multi-child household will be done with substantial error. This must be regarded as a limitation in this study which could be overcome only through data designed to measure parents' perceptions of and possible reactions to differences in children's innate ability.

Because children in the 1981–82 reinterview will be in various development states, we will have to assess their progress relative to peers, that is, those of the same age and sex. We know from work with these data that there are very distinct age- and sex-of-child specific patterns to time use and child development. The empirical work examines the relation between parental time and other family variables in 1975–76 and seven different teacher ratings of the child's cognitive skills relative to age peers in 1982. The main finding is that about 25 to 40 percent of the variation in the individual teacher ratings can be explained by three sets of variables: mother's child-care time and market work in 1975–76, number of male or female siblings present in various age intervals in 1975–76, and family variables, particularly total family income. Mother's education and age and sex of the child had minor statistical impacts in the teacher performance ratings, although

point estimates of the elasticity with respect to education were often substantial. Father's time as an explanatory variable had little relationship, and missing data for fathers reduced our small sample even further. For these reasons, father's time has been excluded from this analysis.

The measures of cognitive skill were teacher responses to the following descriptions of the child:

1) He/she is unable to concentrate: 1 = most of the time; 3 = some of the time; 5 = none of the time (CONCENTRATE).

2) He/she does poorly in school: 1 = most of the time; 3 = some of the time; 5 = none of the time (DOES POORLY).

3) How well did the child comprehend class discussions? 1 = very poorly; 2 = listened but rarely understood; 3 = listened and usually understood; 4 = showed outstanding comprehension (COMPREHEND).

4) How well did he/she retain information? 1 = very poorly; 2 = retained simple ideas and procedures if repeated; 3 = good immediate recall, delayed recall not always good; 4 = immediate and delayed recall; 5 = superior memory (RETAIN INFO).

5) How well does the child grasp new classroom material? 1 = very poorly; 2 = able to grasp simple material only; 3 = always able to grasp simple material, varied ability to grasp more advanced concepts; 4 = usually able to grasp simple and more advanced concepts; 5 = outstanding learner, readily grasped new principles (GRASP MATERIAL).

6) How well did the child work independently? 1 = not at all well; 2 = not very well; 3 = often able to work independently with the straightforward material; 4 = usually able to work independently unless faced with difficult or novel situations; 5 = always worked independently, sought help only as a last resort (WORK INDEPENDENTLY).

7) How well did he/she do in language arts or English last year? 1 = very poorly to 7 = outstanding (sic) (LANGUAGE ARTS).

Estimation consists of predicting teacher ratings in 1982 as a function of age and sex of the child in 1982 and as a function of siblings, parental characteristics, and mother's time allocation in 1975–76. From the theoretical model in Section I it would be desirable to specify both children choices (number and care per child) and career choices (work history and market hours) as functions of a parsimonious set of variables indexing basic market skills and interest. This would have involved a comprehensive statistical model of fertility, market work, and cognitive achievement of the children. Here we have proceeded by treating the educational, marriage, and fertility decisions as exogenous. There is a large extant empirical literature on these issues. Our main interest is in the effect of fertility and market work on child school achievement.

A market wage was estimated for the full sample, allowing for sample selection in the participation equation. The purpose is to determine whether market-earnings capacity is a measure of skills more generally and could therefore explain a parent's success in child rearing.[3] The teacher-rating equations are presented in Table 1. Column 8 of Table 1 is based on the idea that each of the seven teacher ratings is an indicator of an underlying, unobserved variable, school performance. The estimation postulates that a linear combination of the individual ratings is explained by the exogenous variables. This type of multiple-indicator model is discussed in the literature on latent variables (Dennis Aigner and Arthur Goldberger, 1977) and is estimated by maximum likelihood methods in a model proposed for multiple indicators and multiple causes (MIMC) of a single latent variable (Karl Jöreskog and Arthur Goldberger, 1975).

The effect of siblings in 1975 on teacher ratings in 1982 is primarily negative, particularly for siblings in the same (preschool) age range. The number of male siblings in the age ranges 0–2, 3–4, and 5–12 seems to have a *larger* depressing effect on subsequent teacher ratings than the number of female siblings. To illustrate, if one had a brother age 0–2 and another age 3–4, it would lower

[3]The estimation of the market wage and participation equations are given in a longer working paper.

TABLE 1—1982 TEACHER RATINGS OF GRADE-SCHOOLERS' COGNITIVE SKILLS
(Coefficient/Std. Error/Elasticity at Mean)

Independent Variables	Concentrate (1)	Does Poorly (2)	Comprehend (3)	Retain Info (4)
Siblings (1975)				
Males 0–2	−.13/.39/−.013	−.80/.37/−.065	−.16/.16/−.017	−.49/.24/−.045
Females 0–2	.37/.39/ .042	−.35/.37/−.033	−.12/.16/−.014	−.13/.24/−.014
Males 3–4	−.91/.43/−.091	−.68/.41/−.059	−.42/.18/−.045	−.61/.27/−.058
Females 3–4	.44/.49/ .029	−.08/.47/−.004	−.49/.20/−.034	−.35/.30/−.022
Males 5–12	.26/.30/ .036	−.26/.29/−.031	−.33/.13/−.050	−.29/.19/−.040
Females 5–12	.20/.26/ .035	−.05/.25/−.008	−.12/.11/−.023	−.16/.16/−.026
Males 13–17	−.23/.58/−.003	.54/.55/ .007	.21/.24/ .004	.42/.36/ .006
Females 13–17	−.41/.56/−.004	−.43/.53/−.010	−.14/.23/−.004	−.48/.34/−.012
Child Variables				
Age	.04/.10/ .099	−.03/.09/−.077	.07/.04/ .197	.02/.06/ .058
Sex (1 = Female)	−.09/.40/−.014	.34/.39/ .048	−.08/.17/−.014	−.01/.25/−.012
Family Variables				
Family Income	.13/.06/ .462	.17/.06/ .502	.01/.03/ .058	.07/.04/ .223
Mother's Ed	−.03/.11/−.134	−.01/.10/−.039	.04/.04/ .177	.03/.07/ .130
Predicted Wage	.08/.19/ .144	.09/.18/ .143	−.02/.08/−.031	.01/.12/ .024
Mother's Time (1975)				
Care Time	.0003/.0005/ .042	.0005/.0005/ .068	−.0000/.0002/−.003	.0001/.0003/ .013
Helping, Teaching	.0005/.0008/ .001	.0048/.0077/ .009	.0054/.0034/ .012	.0112/.0050/ .022
Talking, Reading	.0008/.0023/ .037	.0034/.0022/ .051	.0014/.0009/ .027	.0016/.0014/ .126
Other Care	.0006/.0008/ .028	−.0003/.0007/−.011	.0001/.0003/ .006	.0001/.0005/ .005
Market Work	−.0002/.0002/−.028	−.0004/.0002/−.070	−.0002/.0001/−.045	−.0002/.0001/−.043
F-test on set of mother's time variables	a	a	a	a
R^2/S.E.E.	.257/1.24	.381/1.18	.369/.518	.350/.771
N	77	77	77	77

Independent Variables	Grasp Material (5)	Work Independently (6)	Language Arts (7)	Lisrel Model[b] (8)
Siblings (1975)				
Males 0–2	−.52/.25/−.048	−.48/.37/−.042	−.31/.43/−.022	−.17/.06
Females 0–2	−.07/.24/−.007	−.07/.36/−.007	−.02/.42/−.001	−.04/.06
Males 3–4	−.57/.28/−.054	−.87/.41/−.080	−1.22/.48/−.088	−.28/.07
Females 3–4	−.42/.31/−.026	−.15/.46/−.009	−.92/.54/−.045	−.14/.06
Males 5–12	−.34/.19/−.045	−.13/.28/−.020	−.50/.33/−.051	−.16/.06
Females 5–12	−.06/.17/−.010	−.22/.24/−.044	−.31/.28/−.040	−.08/.06
Males 13–17	.30/.37/ .004	.19/.55/ .003	.62/.64/ .007	.08/.06
Females 13–17	−.26/.36/−.007	−.23/.52/−.007	−.33/.61/−.006	−.08/.06
Child Variables				
Age	.04/.06/ .191	.08/.09/ .234	.01/.10/ .015	.05/.06
Sex (1 = Female)	−.18/.26/−.028	.23/.38/.425	.56/.44/ .065	.03/.06
Family Variables				
Family Income	.08/.04/ .277	.15/.06/ .570	.18/.07/ .445	.24/.07
Mother's Ed	.05/.07/ .181	−.13/.10/−.572	−.08/.12/−.229	.05/.06
Predicted Wage	.05/.12/ .090	.35/.18/ .685	.10/.21/ .126	
Mother's Time (1975)				
Care Time	.0002/.0003/ .030	.0005/.0005/ .089	−.0004/.0005/−.054	.007/.062
Helping, Teaching	−.0005/.0052/−.001	.0011/.0076/ .003	.0098/.0089/ .015	.095/.063
Talking, Reading	.0018/.0014/ .029	.0019/.0021/ .038	.0030/.0025/ .037	.111/.063
Other Care	.0009/.0005/ .042	.0001/.0008/ .008	.0012/.0009/ .045	.087/.062
Market Work	−.0002/.0001/−.032	−.0001/.0002/−.014	−.0005/.00025/−.070	−.212/.056
F-test on set of mother's time variables	a		a	
R^2/S.E.E.	378/.795	.302/1.1	.416/1.36	.584[c]
N	77	77	77	77

[a] = .10 or less.

[b] The weight or factor loadings on the 7 indicators of school performance given by the teachers were: 1.00, 1.17, 1.08, 1.27, 1.27, 1.15, 1.30. See Jöreskog and Sörbom, 1975, for a discussion.

[c] The adjusted goodness of fit index is given for (8). This is $G \equiv 1-[\kappa(\kappa+1)2d][\mathrm{tr}(\hat{\Sigma}^{-1}S-I)^2/\mathrm{tr}(\hat{\Sigma}^{-1}S)^2]$, where d is the degrees of freedom in the model, κ is the number of variables, S is the sample covariance matrix, and $\hat{\Sigma}$ is the covariance matrix implied by the estimated model ($0 \leq G \leq 1$).

the predicted average teacher rating by more than one point. (The standard deviations of the seven teacher ratings average 1.1.) In column 8 we can see that for this illustration the latent school performance variable would fall by .91 standard deviations.[4] The strength of the negative effect of closely age-spaced siblings was not anticipated but seems to be a pattern worth noting for future research. The presence of older siblings (13- to 17-year-old males or females) appears to have no clearly adverse effect on cognitive development and suggests that (wide) child spacing may lead to larger per capita inputs to children.

Under the hypothesis advanced by Robert Zajonc and Gregory Markus (1975), the cognitive skills of a preschooler are enhanced by exposure to individuals with higher cognitive skills. If so, having older (and therefore more skilled) siblings improves the preschooler's rate of cognitive development. Under the quantity-quality hypothesis of economic demography (Robert Willis, 1973), the effect of number of children can be thought of as raising the cost of acquiring high levels of child quality, and both the time and money price of achieving higher quality are increased as the number of children within a narrower-age range increases. Under the PPP or the selfish-gene model, additional children can be seen as reflecting benefits to the parents at the expense of the existing children.

These results on child spacing can also be interpreted in the context of our framework in Section I. Greater time intervals between child births permit greater inputs to each preschooler. If children are spaced apart very widely, this could lead to a very limited full-time work history and would be represented by high levels of c throughout the planning period. This would be associated with lower-labor market income ($l\alpha S$) by reason of fewer work hours as well as by a smaller stock of market skills from on-the-job training.

The age of a child bears some statistical relation to the teacher's evaluation. This is surprising since the teacher ratings are presumably relative to other children in the same grade. The average elasticity of the teacher-rating variables with respect to age is quite large (averaging .09 as noted in column 8), even though the precision of the underlying parameter estimates is low. There is a tendency for schoolgirls to be evaluated somewhat higher than schoolboys, but this difference is substantial only in the case of language arts. Even here, the standard error is quite large relative to the coefficient.

Family income generally has a positive and significant relation to teacher ratings. This is interesting because these are within school district comparisons so that family resources are less likely to be simple indicators of public school quality. Elasticities at the mean range from .06 to .57 and the column 8 coefficient is .24. The mother's education has no clear statistical influence on the cognitive measures. While the mother's education will influence the number of children, family income, and possibly child spacing—and thereby indirectly influence cognitive skills—it does not seem to have a major statistical influence beyond these routes.

In contrast to the low-statistical power of mother's education on the teacher ratings, the elasticity of the teacher ratings with respect to mother's education is important in some cases. The mother's market wage rate has a positive relationship to the individual teacher ratings, but only in one case is it statistically significant. In the MIMC model, where predicted wage is not included, mother's education has a positive but not statistically significant effect on school performance. Combined with the rather weak effect of education, this suggests that there are different spheres of competence: women who are more productive in the market are not necessarily equally advantaged in their nonmarket skills, and these differences explain market versus home career choices of the type highlighted in the model of Section I.

[4] The standard deviation of Males 0–2 is .50 and Males 3–4 is .48. Since the coefficients in (8) are the effects of movements in the standard deviation of the latent school performance for standard-deviation changes in the explanatory variables, we have the calculation as $(.17/.50)+(.28/.48)=.91$.

The 1975 mother's home-time variables have a zero or positive effect on subsequent teacher ratings in 1981 in 25 of 28 cases and the effect of market time on the teacher ratings is negative in all seven cases. In only one of seven teacher ratings is an F-test on the hypothesis of no effect for the set of mother's market and nonmarket time variables not rejected at a 10 percent level. Not only does mother's market work time have a negative effect on all seven of the teacher ratings, but the effect is statistically significant at the 10 percent level for four of the seven teacher ratings. In our summary model in (8), mother's market work has a significant, negative effect on school performance.

For a typical coefficient on the weekly minutes of market work of about $-.0003$, increasing market work from the average of about 10 hours per week (620 minutes) to full-time work of 40 hours would reduce the teacher's ratings by about one-half point.[5] Thus, while market time and home time appear to be important, they are certainly no more important than number of siblings in nearby age ranges. Moreover, women who work are likely to have fewer children, therefore per child time is not generally less for working mothers. In samples of this sort one would not expect the simple correlation between measures of work and children's school performance to be significantly negative. If one ignores measurement error issues and takes the regression coefficients in Table 1 literally, then, for example, market work would increase language arts performance as long as the hourly wage rate exceeded \$2.78. At this wage rate the benefits of higher family income offset the negative effects of market work hours.

If we tentatively accept the negative relation between market time and children's school performance, what other interpretations should be given? One is that suggested by a theory of energy allocation (Becker, 1985). Another interpretation along these lines is that children's development depends on the *timing of time* devoted to child care. Attention dependent of child-initiated requests may be more important for development, and simple absence of a parent via market work reduces the opportunities for this type of interaction, even if total child-care time as measured by our child-care variables is the same. A question that cannot be answered by these data is whether specific types of child care can be effective substitutes for parents' own time, or, for example, whether special child-care arrangements in the workplace are a substitute for reduced market work. A third interpretation suggested by the earlier discussion of self-indulgent parents is that those mothers (and their spouses) choosing more market work simply have a greater relative preference for own market goods. (The implications of a smaller γ can be easily shown in Figure 1.) In any event the results appear challenging in light of recent evidence on the importance of full-time market work as the vehicle for career advancement of women.

III. Summary

At this point we do have a preliminary picture of the relation between parental care and school performance: Mother's time does seem to matter, though measurement error reduces the apparent strength of the relationship. The negative effect of market time on the school attainment and future earnings of offspring has been shown in recent research (Martha Hill and Greg Duncan, 1987). Siblings, particularly those closely age-spaced to the child, appear to lower subsequent cognitive performance as measured by teacher ratings, and the adverse effect of siblings seems to be greater if the siblings are male. Since the presence of older siblings does not appear to have any clearly negative impact on subsequent cognitive skills, this suggests a benefit to child spacing, although the implied time intervals are quite wide and, as indicated by the theoretical model and recent empirical studies, this would imply substantial costs in terms of lifetime market earnings.

Family income has a statistically significant impact on teacher ratings (in 6 of 7 cases), and this is shown in the summary equation with school performance as a latent

[5] In equation (8) an increase from 10 to 40 hours is an increase of 2.1 standard deviations in market work. This would imply a $2.1 \times .212 = .45$ standard-deviation reduction in school performance.

variable indicated by the seven teacher ratings. In contrast to findings of others (Linda Datcher-Loury, 1986; Murnane et al., 1981), mother's education and market wage have a weak statistical relation to the teacher ratings. Mother's education leads to fewer children and greater care time to each child, thereby indirectly influencing the "per capita" level of time and the resulting cognitive skills of the child. Another interpretation is that mother's education represents an unmeasured greater preference for own consumption (parental prodigal effects), which is reflected in a greater life-cycle market skill investment as suggested by the theoretical model of Section II.

REFERENCES

Aigner, Dennis J. and Goldberger, Arthur S. eds., *Latent Variables in Socioeconomic Models*, Amsterdam: North-Holland, 1977.

Becker, Gary S., "Human Capital, Effort and the Sexual Division of Labor," *Journal of Labor Economics*, January 1985, *3*, S33–S58.

______, *A Treatise on the Family*, Cambridge: Harvard University Press, 1981.

______ **and Tomes, Nigel,** "An Equilibrium Theory of the Distribution of income and Intergenerational Mobility," *Journal of Political Economy*, December 1979, *87*, 1153–89.

Bergstrom, Theodore, "A Fresh Look at the Rotten Kid Theorem and Other Household Mysteries," *Journal of Political Economy*, forthcoming 1988.

Corcoran, Mary, Duncan, Greg J. and Ponza, Michael, "A Longitudinal Analysis of White Women's Wages," *Journal of Human Resources*, Fall 1983, *18*, 497–520.

Datcher-Loury, Linda, "Effects of Mother's Home Time on Children's Schooling," unpublished manuscript, 1986.

Dawkins, Richard, *The Selfish Gene*, New York: Oxford University Press, 1976.

Gustafsson, Siv and Jacobsson, Roger, "Trends in Female Labor Force Participation," *Journal of Labor Economics*, January 1985, *3*, S256–S274.

Hess, Robert D. and Holloway, Susan D., "Family and School as Educational Institutions," in R. D. Parke, ed., *The Family: Review of Child Development Research*, Vol. 7, Chicago: The University of Chicago Press, 1984.

Hill, Martha S. and Juster, F. Thomas, "Constraints and Complementarities in Time Use," in F. T. Juster and F. P. Stafford, eds., *Time, Goods, and Well-Being*, Ann Arbor: Institute for Social Research, University of Michigan, 1985.

______ **and Duncan, Greg J.,** "Parental Family Income and the Socioeconomic Attainment of Children," *Social Science Research*, March 1987, *16*, 39–73.

Jöreskog, Karl G. and Goldberger, Arthur S., "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable," *Journal of the American Statistical Association*, September 1975, *70*, 631–39.

______ **and Söbom, Dag,** "Statistical Models and Methods for the Analysis of Longitudinal Data," in D. J. Aigner and A. S. Goldberger, eds., *Latent Variables in Socioeconomic Models*, Amsterdam: North-Holland, 1977.

Juster, F. Thomas and Stafford, Frank P., eds., *Time, Goods, and Well-Being*, Ann Arbor: Institute for Social Research, University of Michigan, 1985.

Murnane, Richard J., Maynard, Rebecca A. and Ohls, James C., "Home Resources and Children's Achievement," *Review of Economics and Statistics*, August 1981, *63*, 369–77.

Ryder, Harl, Stafford, Frank P. and Stephan, Paula E., "Labor, Leisure, and Training over the Life Cycle," *International Economic Review*, October 1976, *17*, 651–74.

Willis, Robert J., "A New Approach to the Economic Theory of Fertility Behavior," *Journal of Political Economy*, March-April 1973, *81*, S14–64.

Wohlwill, Joachim F., "Cognitive Development in Childhood," in O. G. Brim and J. Kagen, eds., *Constancy and Change in Human Development*, Cambridge: Harvard University Press, 1980.

Zajonc, Robert and Markus, Gregory B., "Birth Order and Intellectual Development," *Psychological Review*, January 1975, pp. 74–88.

Do Biases in Probability Judgment Matter in Markets? Experimental Evidence

By COLIN F. CAMERER*

Microeconomic theory typically concerns exchange between individuals or firms in a market setting. To make predictions precise, individuals are usually assumed to use the laws of probability in structuring and revising beliefs about uncertainties. Recent evidence, mostly gathered by psychologists, suggests probability theories might be inadequate *descriptive* models of *individual* choice. (See the books edited by Daniel Kahneman et al., 1982a, and by Hal Arkes and Kenneth Hammond, 1986.)

Of course, individual violations of normative theories of judgment or choice may be corrected by experience and incentives in markets, thus producing market outcomes which are consistent with the individual-rationality assumption even if that assumption is wrong for most agents. Whether judgment and choice violations matter in markets is a question that begs for empirical analysis.

In this paper I use experimental markets to address this issue (see also Rong Duh and Shyam Sunder, 1986; and Vernon Smith, 1982, for an overview). In these markets, traders are paid dividends for holding a one-period asset. The amount of the dividend depends upon which of two states occurred. Traders know the prior probabilities of the states, and a sample of likelihood information about which state occurred. The setting is designed so that prices and allocations will reveal whether traders use Bayes' rule to integrate the prior and the sample information, or whether they judge the likelihood of each state by the "representativeness" of the sample to the state (Amos Tversky and Kahneman, 1982b). (Several other non-Bayesian psychological theories can be tested, too.)

Evidence of judgment bias reported by psychologists poses an implicit challenge to economic theory based on rationality. Sometimes that challenge is made explicit, as when Kenneth Arrow suggested that use of the representativeness heuristic "typifies very precisely the excessive reaction to current information which seems to characterize all the securities and futures markets" (1982, p. 5). Others have warned that judgment biases will affect the judgments of well-trained experts who make societal decisions (about the risk of low-probability hazards, for instance, see Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein, 1976).

Assertions as bold as Arrow's are extremely rare, because the faith that individual irrationality will not affect markets is a strong part of the "oral tradition" in economics. This faith is often defended with Milton Friedman's (1953) famous claim that theories with false assumptions (such as strong assumptions of individual rationality) might still predict market behavior well (see Mark Blaug, 1980, pp. 104–14, for a cogent discussion). Besides that "*F*-twist," there is a standard list of arguments used to defend economic theories from the criticism that people are not rational. (Counterarguments are given in parentheses.)

1) In markets, agents have enough financial incentive, and experience, to avoid mistakes. (Incentives and experience were provided in David Grether's 1980 experiments on the representativeness heuristic. See

*Department of Decision Sciences, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. I thank Mike Chernew, Marc Knez, Peter Knez, and Lisabeth Miller for research assistance. Helpful comments have been received from three anonymous referees, Greg Fischer, Len Green, David Grether, Dan Kahneman, John Kagel, Peter Knez, George Loewenstein, Charles Plott, Paul Slovic, Shyam Sunder, Richard Thaler, Keith Weigelt, and especially Howard Kunreuther. This research was funded by the Wharton Risk and Decision Processes Center, the Alfred Sloan Foundation grant no. 8551, and the National Science Foundation grant no. SES-8510758.

also Charles Plott and Louis Wilde, 1982, p. 97.)

2) Random mistakes of individuals will cancel out. (The biases found by psychologists are generally *systematic*—most people err in the same direction.)

3) Only a small number of rational agents are needed to make market outcomes rational, if those agents have access to enough capital or factors of production. (Institutional constraints may prevent those agents from making markets rational; see Thomas Russell and Richard Thaler, 1985.)

4) Agents who are less rational may learn *implicitly* from the actions of more rational agents. (This argument requires that "more rational" agents are identifiable, perhaps by their more vigorous trading.)

5) Agents who are less rational may learn *explicitly* from more rational agents by buying advice or information. (Institutional constraints, and the well-known problems of adverse selection and moral hazard, may limit the extent of information markets.)

6) Agents who are less rational may be driven from the market by bankruptcy, either by natural forces or at the hands of more rational competitors. (A new supply of agents who are less rational, or inexperienced, may be constantly entering the market.)

Most of these arguments, though not all of them, are put to the test in the market experiments described below. Subjects trade for up to 7 hours, observing nearly 100 realizations of the state variable, and every trade earns them a (small) dollar profit or loss (argument 1). The representativeness heuristic is systematic in direction (argument 2). Subjects trade with one another in a "double-oral" auction with no constraints on bidding or offering activity (argument 3), so they can learn implicitly from others' trading behavior (argument 4).

Many of the standard arguments are *not* tested in the experiments: There is no explicit market for advice (argument 5); subjects cannot sell short (argument 3); and bankruptcy is unlikely, though conceivable (argument 6). The first two arguments are being tested in further work. Even with these limits, the market experiments provide a greater combination of incentives, experience, and learning opportunity than in previous judgment experiments.

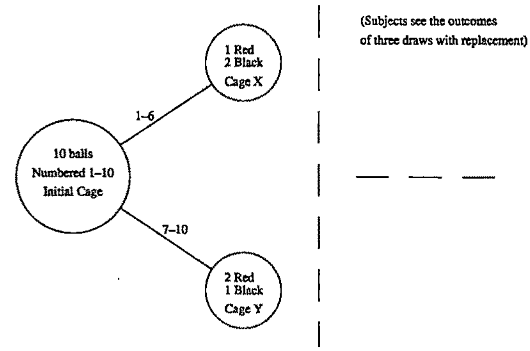


FIGURE 1

I. Experimental Design

In the experiments, each of 8 or 10 traders is endowed with two assets that live one period and pay a liquidating state-dependent dividend.

A. *State Probabilities*

The state is represented by which one of two bingo cages (X or Y) is chosen (Figure 1). A third bingo cage containing 10 balls is used to determine whether cage X or cage Y has been chosen. The X cage contains 1 red and 2 black balls. The Y cage contains 2 red balls and 1 black one. The prior probabilities of X and Y are .6 and .4.[1] Figure 1 is shown on a blackboard for all subjects to see, throughout the experiment.

After either X or Y is chosen (but *not* announced), a sample of three balls is drawn from the chosen cage, with replacement, and the sample is announced before trading begins. Since the cages X and Y contain different populations of balls, which are known to traders, they can use Bayes' rule to calculate $P(X/\text{sample})$ from the prior $P(X)$ and

[1] Unequal priors were chosen because priors of .5 and .5 might have made it too easy for subjects to intuit the Bayesian posteriors. Experiments with equal priors are a natural direction for future work.

TABLE 1—BAYESIAN EXPECTED DIVIDEND VALUES

				Bayesian Posterior $P(X/\text{sample})$			
				.923	.750	.429	.158
				Bayesian Expected Values[a]			
	No. of Traders	Dividend		No. of Reds			
Type	(Experiment No.)	X	Y	0	1	2	3
I	5 (1,3–5,11x–15xh) 4 (2,9r,10h)	500	200	477	425	329	247
II	5 (1,3–5,11x–15xh) 4 (2,9r,10h)	350	650	373	425	521	603
I	5 (6–7) 4 (8,12x)	525	225	502	450	354	272
II	5 (6–7) 4 (8,12x)	180	480	203	255	351	433

Note: All dividends were actually 80 francs higher, for both types of traders and in both states, in experienced subjects experiments 11x, 13x–15xh. (Therefore, all Bayesian expected values are 80 francs higher, too.) In all analyses prices are adjusted for this 80-franc difference.

[a] In francs.

the likelihood functions $P(\text{sample}/X)$ and $P(\text{sample}/Y)$ (which are determined by the cage contents). The top line of Table 1 gives the Bayesian posteriors for all three-ball samples. The possible samples are characterized by the number of reds only, since the order of draws should not matter and the data suggest the order did not matter to subjects. (In some experiments, like John Hey's 1982 experiments on price search, order does seem to matter. Subjects were paid in his 1987 experiments and order still mattered.)

B. *Market Procedure*

Subjects were undergraduate men, and some women, recruited from quantitative methods and economics classes at the Wharton School. These students have all taken statistics and economics courses. Experiments 1 to 10 used subjects who had not been in any previous market experiments. Five experiments used "experienced" subjects who had been in experiments 1 to 10; these experiments are numbered 11x to 15xh (the "x" reminds the reader that subjects were experienced). Experiments were conducted in one 3-hour session (experiments 1 and 2, 6, 9 and 10, 11x to 15xh) or two 2-hour sessions held on consecutive evenings (experiments 3 to 5, 7 and 8).

All trading and earnings are in terms of francs, which are converted to dollars at the end of the experiment at a rate of $.001 dollars per franc ($.0015 in experiment 1).[2] Traders are endowed with 10,000 francs and two certificates in each trading period, and 10,000 francs is subtracted from their total francs at the end of each period. In some experiments a known fixed cost (around 5,000 francs) was subtracted from their total earnings at the end of the experiment.

Traders voluntarily exchange assets in a "double-oral auction": Buyers shout out bids at which they will buy, sellers shout out offers at which they will sell. Bids must top outstanding bids and offers must undercut outstanding offers. A matching bid and offer is a trade, which erases all previous bids and offers. All bids, offers, and trades in a period are recorded by the experimenter on a transparency visible to subjects. (No history of previous periods of trading is posted.) Trading periods last 4 minutes in 10-subject experiments, 3 minutes in 8-subject experiments.

At the end of each trading period the state (X or Y) is announced and traders calculate

[2] In practice, using francs makes traders more precise in their trading than they would be with dollars, for example, traders routinely haggle over 5-franc differences between bids and offers, which represent half a penny. Francs may also alleviate competition among traders for relative status in dollar earnings, because traders' dollar conversion rates (while identical) are privately known.

their profits. Dollar profits are given by

$$(1)\ PROFITS = X\left[E_f - R_f + \sum_{i=1}^{x_s} 0_i - \sum_{j=1}^{x_b} B_j + D(S) \times (E_c - x_s + x_b) - F\right],$$

where X = dollar-per-franc conversion rate,
E_f = initial endowment in francs,
R_f = amount of francs repaid at period-end,
E_c = initial endowment in certificates,
x_s = number of certificates sold,
0_i = selling price of ith certificate sold,
x_b = number of certificates bought,
B_j = purchase price of jth certificate bought,
$D(S)$ = dividends per certificate in state S,
F = fixed cost per experiment in francs.

Traders may not sell short (that is, $E_c - x_s + x_b$ cannot be negative), and net francs on hand ($E_f + \Sigma 0_i - \Sigma B_j$) cannot be negative.

C. *Market Equilibrium*

Assuming risk neutrality, traders' reservation prices for assets are expected values. (If they are not risk neutral, their reservation prices are certainty equivalents.) Since each trader's endowment of francs is large enough to buy virtually the entire market supply of assets, and the supply is fixed (by the initial endowment, and the short-selling restriction), there is excess demand at any price less than the highest expected value. Thus, in competitive equilibrium, prices should be bid up to the largest expected value of any trader. One irrational trader who pays too much can therefore create a market price that is too high. The empirical question is whether such traders exist, and whether the experience and financial discipline of a market makes them more rational over the course of an experiment.

Of course, the double-oral auction is not Walrasian, so there is no theoretical assurance that competitive equilibrium will result. However, simple models of the double-oral auction as a dynamic game with incomplete information are beginning to establish the theoretical tendency of double-oral auctions to converge to competitive equilibrium (Daniel Friedman, 1984; Robert Wilson, 1985; see David Easley and John Ledyard, 1986). The empirical tendency to converge is well-established (for example, Smith, 1982), even in designs meant to inhibit convergence (Smith and Arlington Williams, in press).

D. *Competing Theories*

In each experiment, traders are randomly assigned to either of two "types," which differ in the dividends they receive in the two states X and Y (see Table 1). The dividends are chosen so that competing theories predict different patterns of prices and allocations (see Table 2). Each theory will now be described briefly.

Bayesian. If traders use Bayes' rule to calculate posterior probabilities given the sample data, prices should converge to the Bayesian expected values given in Table 2, assuming risk neutrality. (Tests and controls for risk neutrality are described below.) In the experiments described by the top panel of Table 2, for instance, type I traders should pay up to 477 if the sample is 0 reds, 425 if 1 red, 329 if 2 reds, and 247 if 3 reds. Type II traders should pay up to 373, 425, 521, and 603, respectively. Therefore, if the sample is 0 reds, then type I traders should buy from type II traders at a price of 477. If the sample is 2 or 3 reds, the type II traders should buy all the units, at prices of 521 or 603, respectively. If the sample is 1 red, then type I and type II traders both have a Bayesian expected value of 425 francs, so we expect half the units will be held by each of the two types of traders. (Trades might take place because of uncontrolled differences in risk tastes, but units are still equally as likely to end up in the hands of type I and type II traders.) In experiments 6 to 8 and 12x, dividends were chosen so that the Bayesian expected values of the type I and type II

TABLE 2—PRICE AND ALLOCATION PREDICTIONS OF COMPETING THEORIES

	Predictions Expressed as: Price P (Type Holding Assets)			
		Number of Reds in Sample		
Theory	0 Reds	1 Red	2 Reds	3 Reds
Experiments 1–5, $9r$–$11x$, $13x$–$15xh$				
Bayesian	477 (I)	425 (I, II)	521 (II)	603 (II)
Exact Representativeness	477 (I)	$P > 425$ (I)	$P > 521$ (II)	603 (II)
Conservatism	$P < 477$ (I)	$P > 425$ (II)	$P < 521$ (II)	$P < 603$ (II)
Overreaction	$P > 477$ (I)	$P > 425$ (I)	$P > 521$ (II)	$P > 603$ (II)
Base-Rate Ignorance	467 (I)	450 (II)	550 (II)	617 (II)
Experiments 6–8, $12x$				
Bayesian	502 (I)	450 (I)	354 (I)	433 (II)
Exact Representativeness	502 (I)	$P > 450$ (I)	$P > 354$ (II)	433 (II)
Conservatism	$P < 502$ (I)	$P < 450$ (I)	$P > 354$ (I)	$P < 433$ (II)
Overreaction	$P > 502$ (I)	$P > 450$ (I)	$P > 354$ (II)	$P > 433$ (II)
Base-Rate Ignorance	492 (I)	425 (II)	380 (II)	447 (II)

traders were (nearly) equal when a 2-red sample was drawn.

Exact Representativeness. If subjects take the representativeness of the sample to the cage contents as a psychological index of the cage's likelihood, non-Bayesian expected values might result. Representativeness is a vague notion, but we can distinguish some precise variants of it. For instance, subjects might think $P(X/\text{sample}) = 1$, if the sample resembles the X-cage contents more closely than the Y-cage contents. Or they might think $P(X/\text{sample}) = 1$, if the sample exactly matches the X-cage contents. These extreme hypotheses are clearly ruled out by the data presented below.

More reasonably, subjects may be intuitively Bayesian for most samples, but overestimate a cage's likelihood when a sample resembles the cage *exactly.* This "exact representativeness" theory predicts that subjects will judge $P(X/1 \text{ red})$ to be greater than the Bayesian posterior .75 because a 1-red sample exactly matches the X-cage's contents. Similarly, $P(Y/2 \text{ red})$ will be judged to be greater than .57; other probabilities will be Bayesian. Of course, there are other possible interpretations but since they are either imprecise or clearly incorrect, only exact representativeness will be considered carefully.

Under exact representativeness, prices will be higher than Bayesian in 1- and 2-red periods (as shown in Table 2) and type I traders will hold units in 1-red periods. (Recall that the Bayesian theory predicts types I and II are equally likely to hold units in 1-red periods.)

Base-Rate Ignorance. If subjects judge $P(\text{state}/\text{sample})$ by the representativeness of samples to states, their judgments may ignore differences in the prior probabilities (or "base rates") of states (Tversky and Kahneman, 1982b). In our setting it is difficult to integrate this aspect of representativeness with other aspects, like the psychological power of exact representativeness, because the two aspects often work in opposite directions. In 1-red samples, for instance, exact representativeness predicts $P(X/1 \text{ red})$ will be overestimated, while ignorance of the higher base rate of X implies $P(X/1 \text{ red})$ will be underestimated. Since predictions of a theory that integrates representativeness with base-rate ignorance are ambiguous, I define base-rate ignorance as using Bayes' rule with erroneous priors $P(X) = P(Y) = .5$. Predictions of this theory are shown in Table 2.

Of course, ignoring base rates completely is rather implausible. For example, in an experiment with a prior probability of .001, it seems unlikely that subjects will act as if the prior is .5. If priors are simply underweighted, but not ignored, the data will show some statistical support for the complete base-rate ignorance theory. The theory

should be considered an extreme benchmark that helps us judge whether priors are underweighted at all.

Conservatism. Subjects may be "conservative" in adjusting prior probabilities for sample evidence (for example, Ward Edwards, 1968).

Overreaction. Subjects may adjust prior probabilities *too much*, as if overreacting to sample evidence. The overreaction theory makes the same prediction as representativeness in 1- and 2-red periods, but it predicts bias in 0- and 3-red periods where representativeness does not. Note that the conservatism and overreaction theories make exactly opposite predictions. This implies quite a challenge for the Bayesian theory: Prices must be exactly at the Bayesian prediction, or insignificantly different from it, for both theories to be falsified.

II. Results

Fifteen experiments have been conducted —ten with inexperienced subjects, five with experienced subjects—excluding two inconclusive pilot experiments. For the sake of brevity, many details of the analyses are omitted and can be found in working papers available from the author.

There are two kinds of data which distinguish between theories: prices at which trades occurred, and the number of units of the asset that traders held at the end of trading periods.

A. *Trade Prices*

The mean prices across experiments 1 to 8 are summarized by a time-series of 90 percent confidence intervals, shown in Figure 2.[3] The upper (lower) solid line is the upper (lower) end of the confidence interval. Bayesian expected values are shown by dashed lines, and the direction of the exact representativeness prediction is shown by an arrow marked "*R*." Each of the four panels represents a different sample. From left to right, observations within a panel represent data from the first time that sample was drawn, the second time the same sample was drawn, and so forth.

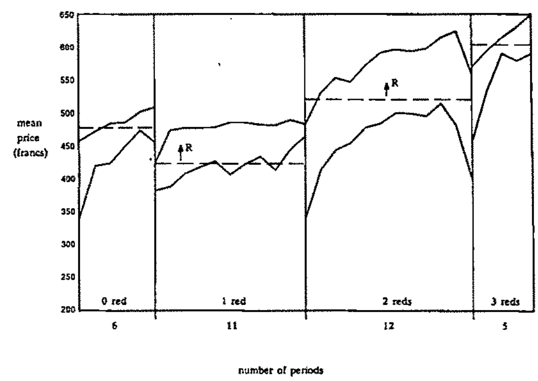


FIGURE 2

Prices converge, from below, toward the Bayesian levels. These data clearly rule out many non-Bayesian theories of probability judgment (like the two extreme brands of representativeness mentioned above). However, prices do not converge exactly to the Bayesian expected values. There is some evidence of exact representativeness, because prices drift above the Bayesian expected values in 1- and 2-red periods. However, the confidence intervals are wide, and the degree of bias is rather small. Indeed, since prices should only converge to Bayesian predictions if the hypotheses of risk neutrality,

[3]Confidence intervals were constructed by first calculating mean prices in each period of each experiment, then separating the time-series of mean prices for each different sample. Data from experiments 6 to 8 were normalized so that the Bayesian predictions in those experiments were the same as in experiments 1 to 5. This yields groups of data such as 8 mean prices from the first 0-red period in each of the 8 experiments numbered 1 to 8. The mean of those means, and its standard error (the standard deviation divided by $8^{1/2}$) are used to calculate the 90 percent confidence interval. A second confidence interval was calculated using mean prices from the second 0-red period in each of the 8 experiments, and so on. Not all experiments have the same number of 0-red periods, so the number of observations in each confidence interval gradually decreases. The procedure was stopped just before there was only one experiment left with an Nth observation of a particular sample.

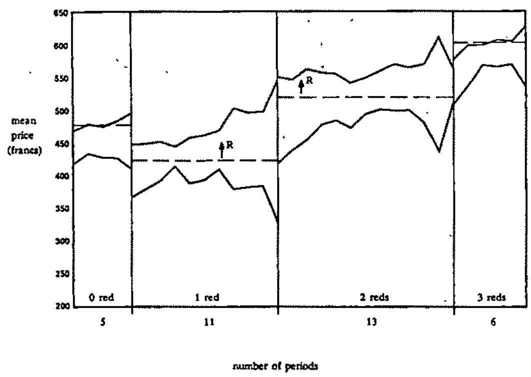


FIGURE 3

competitive equilibrium, and Bayesian updating are all true simultaneously, it is rather remarkable that prices converge as closely to the Bayesian predictions as they do.

Figure 3 shows confidence intervals from experiments with experienced subjects. Prices begin closer to the Bayesian expected value, and have less tendency to drift above it in 1- and 2-red periods. The confidence intervals are also wide, because they summarize a small number of experiments.[4]

We can define bias in prices as a deviation from the Bayesian prediction. If the Bayesian theory is true, biases will be around zero. To conduct statistical tests on price biases, the time-series of prices in each experiment must be independent. Since prices are typically autocorrelated, the equilibrium degree of bias is estimated from a simple partial adaptation model (a first-order autoregression),

$$(2)\quad P_t - P_{\text{Bayes}} = a + b\left(P_{t-1} - P_{\text{Bayes}}\right) + e_t,$$

where P_t is the tth observation of price and P_{Bayes} is the Bayesian prediction. This specification implies that the deviation from equilibrium is reduced by a fraction $1-b$ each trade. If b is close to 1, convergence is very slow; if b is close to 0, convergence is fast. While there is no theoretical rationale for (2), it works well empirically and there is no well-established theory of price convergence which suggests it is wrong.

Call the bias for the tth price B_t; it equals $P_t - P_{\text{Bayes}}$. If we define equilibrium as a bias that does not change each period, we impose $B_t = B_{t-1} = B$ on (2) and get

$$(3)\qquad B = a + bB + e_t.$$

Since $E(e_t) = 0$, a little algebra shows that we can estimate the degree of equilibrium bias B consistently by the estimator $B' = a'/(1-b')$, where a' and b' denote ordinary least squares estimators of a and b in (2). The standard error of B' can be calculated from a Taylor series approximation involving the variances of a' and b' and their covariance.[5]

Regressions were first run separately for each period, effectively allowing a and b to vary each period. The simple specification (2) fit fairly well: The convergence rate b was typically estimated precisely, and residuals were uncorrelated and roughly homoskedastic. An F-test (Jan Kmenta, 1971, p. 373) was used to test whether adjacent periods could be pooled at the 10 percent level. Periods were pooled, starting with the last period, until the F-test was violated.

The estimate B' resulting from the last group of poolable periods in each experiment are shown in Table 3. Also reported is the t-statistic testing the hypothesis that $B = 0$, which is simply B' divided by its (approximated) standard error. Sample sizes are shown in parentheses next to each experiment number. T-statistics marked with asterisks are unreliable because the assumption of normality of residuals was violated at the 1 percent level, by the studentized range

[4]Intervals flare out in Figure 3 when the number of different experiments used to construct them drops steeply and standard errors increase dramatically.

[5]I thank Dave Grether for correcting a mistake in earlier estimates of $V(B')$. The Taylor series approximation of $a'/(1-b')$ around its true value $a/(1-b)$ is $a/(1-b)+(a'-a)/(1-b)+a(b'-b)/(1-b)^2$, plus some higher-order terms. Using this expression to calculate (approximately) $V(a/(1-b))$, or $E[(a'/(1-b')-a/(1-b))^2]$ yields $V(a')/(1-b)^2 + a^2V(b')/(1-b)^4 + 2a\text{COV}(a',b')/(1-b)^3$. Evaluating this expression at a' and b' gives approximations of $V(B')$.

Table 3—Estimates of Bias in Equilibrium Prices, and Tests of the Bayesian Hypothesis Against Competing Hypotheses

Experiment (n)	Bias B	t-Statistic	Significance Levels, Bayesian vs. Conservatism	Overreaction	Base-Rate Ignorance
		0-Red Periods			
Inexperienced subjects					
1 (24)	−28.31	2.31	.01	.99	.06
2 (46)	19.28	10.95	.999	.000	.000
3 (54)	−11.09	−3.67*	.000	.999	.999
4 (57)	4.97	4.56*	.999	.000	.000
5 (10)	−22.35	−6.57	.000	.000	.999
6 (29)	15.36	5.67*	.999	.000	.999
7 (70)	32.44	.65*	.76	.24	.57
8 (7)	−37.90	−2.20	.99	.01	.12
9r (9)	10.01	8.70	.000	.999	.999
10h (10)	−2.54	−1.16*	.12	.88	.999
mean	−2.95		.49	.51	.57
Experienced subjects					
11x (53)	−44.12	12.15	.000	.999	.999
12x (34)	15.17	1.78	.96	.04	.01
13x (8)	76.50	.49	.69	.31	.51
14x (18)	11.61	3.64	.999	.000	.999
15xh (16)	4.92	1.41	.92	.08	.35
mean		2.14		.71	.29.57
		1-Red Periods			
	Bayesian vs.		Exact Representativeness, Overreaction	Conservatism	Base-Rate Ignorance
1 (13)	5.00	2.63*	.005	.005	.999
2 (40)	56.34	7.94*	.000	.000	.000
3 (40)	1.18	.29*	.46	.46	.999
4 (25)	49.81	18.94	.000	.000	.000
5 (37)	31.19	10.23	.000	.000	.000
6 (28)	51.80	9.10	.000	.999	.999
7 (16)	23.12	4.65	.001	.999	.985
8 (57)	93.83	.18	.43	.57	.51
9r (8)	51.15	6.21	.000	.000	.000
10h (50)	54.63	14.12*	.000	.000	.000
mean	39.92		.09	.30	.45
11x (44)	−2.76	−2.08	.98	.98	.999
12x (7)	32.18	3.82	.001	.999	.999
13x (24)	.96	.49	.31	.31	.999
14x (33)	27.88	1.89*	.03	.03	.21
15xh (8)	29.77	3.49	.005	.005	.001
mean	17.61		.27	.47	.64

(continued)

test. Other diagnostic tests and estimates of b are reported in working papers.

Roughly speaking, biases are distributed around zero in 0-, 2-, and 3-red periods. Biases are positive in 1-red periods of every experiment except 11x, generally with large t-statistics. Biases are also positive in 2-red periods with experienced subjects, but not with inexperienced subjects.

The right-hand columns of Table 3 test the hypothesis that prices are Bayesian against each of the competing theories. The tests of the Bayesian theory against exact representativeness, conservatism, and overreaction are one-tailed t-tests of the null hypothesis $B=0$ against one-sided alternative hypotheses (which vary depending upon the theory and the sample). Since the base-rate ignorance theory predicts a point estimate of the bias rather than a direction, the significance level of the Bayesian hypothesis against the base-rate ignorance alternative was esti-

TABLE 3—(CONTINUED)

	Bayesian vs.		2-Red Periods Exact Representativeness, Overreaction	Conservatism	Base-Rate Ignorance
1 (61)	−27.00	−4.84	.999	.000	.999
2 (24)	60.25	.11	.46	.54	.50
3 (22)	99.00	6.27	.000	.999	.000
4 (83)	77.39	7.74	.000	.999	.000
5 (52)	−53.31	6.55	.76	.24	.64
6 (77)	−1.74	−.02*	.51	.51	.52
7 (16)	98.24	18.16	.000	.000	.000
8 (16)	45.45	3.35	.001	.001	.000
9*r* (24)	−17.23	−4.55	.999	.000	.999
10*h* (27)	−7.57	−.95	.67	.33	.999
mean	27.35		.44	.36	.44
11*x* (18)	49.28	7.55	.000	.999	.000
12*x* (8)	20.80	12.16*	.000	.000	.000
13*x* (15)	17.22	14.35	.000	.999	.000
14*x* (11)	22.62	18.26	.000	.999	.000
15*xh* (17)	12.47	.80	.21	.79	.62
mean	24.48		.04	.78	.12

	Bayesian vs.		3-Red Periods Conservatism	Overreaction	Base-Rate Ignorance
1 (40)	2.47	.40	.65	.35	.04
2 (17)	−209.34	−3.51	.000	.999	.85
3 (41)	41.88	4.12*	.999	.000	.000
4 (48)	11.26	.64*	.74	.26	.41
5 (26)	−10.57	.70	.24	.76	.90
6 (32)	31.55	1.29*	.90	.10	.24
7 (22)	20.04	3.24*	.999	.001	.000
8 (28)	−26.46	−.79	.29	.71	.70
9*r* (29)	14.01	.29	.61	.39	.48
10*h* (7)	2.61	.02	.51	.49	.50
mean	−12.23		.59	.41	.41
(2 deleted)		9.64			
11*x* (37)	−22.51	−6.52	.000	.999	.999
12*x* (9)	24.98	.49	.69	.31	.39
13*x* (35)	11.65	.65	.74	.26	.40
14*x* (28)	114.21	.13	.55	.45	.50
15*xh* (9)	−38.00	−3.74	.000	.999	.999
mean	4.70		.40	.60	.66

Notes: * denotes studentized range of residuals greater than the 1 percent level for normality, so standard errors are unreliable. Biases are truncated in calculating means when the equilibrium price implied by the bias estimate is greater than the maximum dividend for the type of trader holding a majority of units (for example, 0-red period, experiment 7).

mated from likelihood ratios.[6] Significance levels were estimated by assuming the t-statistics were normally distributed (a reasonable approximation for most of the sample sizes in Table 3). Levels less than .001 or above .999 are reported as .000 or .999.

The significance levels of tests against most of the alternative theories are roughly 50 percent, suggesting departures from the Bayesian predictions are not systematic. However, the Bayesian theory can be strongly rejected against the alternative of exact representativeness in most 1-red periods and

[6] P(data/Bayesian) and P(data/Base-rate Ignorance) were calculated assuming the estimate B' was normally distributed with standard deviation $s(B')$. Assuming one of the two theories is true, and they are equally likely; Bayes' rule can then be used to calculate P(Bayesian/data).

many 2-red periods. Of course, the statistical significance of a bias is simply a measure of whether it could be due to chance. Whether the biases are economically significant is discussed in the conclusion.

Note that the graphs and the statistical tests seem to tell different stories because the confidence intervals are wide while the t-statistics are large. This simply means that biases are not random in each experiment (hence, the extreme significance levels in Table 3), but the degree of bias varies a lot across experiments (hence, the wide confidence intervals).

In most experiments subjects did not make probability calculations during the experiment (though they were given calculators to record profits). However, in experiment 1 two traders *did* write the correct likelihood ratios $P(X/\text{sample})/P(Y/\text{sample})$ on their profit sheets during the experiment; prices were quite close to Bayesian (for example, 1-red prices were only 5 francs too high). A small number of aggressive Bayesians apparently can make the market price Bayesian, but did not do so very often.

B. *Allocations of Assets*

For most samples, competing theories all predict the same type of trader will hold units. When the theories make the same prediction, they are extremely accurate. In 0-red and 3-red periods, for instance, virtually all of the units are held by the traders with the highest expected dividend type in every experiment.

The theories disagree about allocations in 1-red periods of some experiments and 2-red periods of other experiments. In these experiments, the average fraction of traders holding any units at the end of the period and the average fraction of units held were calculated for dividend types I and II. These data are shown in Table 4.

In the 1-red periods, the Bayesian theory predicts type I and type II traders are equally likely to hold units (since their expected values are equal, at 425). Exact representativeness predicts units will be held by type I's.

Across all experiments with inexperienced subjects, type I's hold 78 percent of the units. This fraction is quite stable across experiments, and is about the same in early periods (the first half of the periods) and late periods. With experienced subjects, about 90 percent of the units are held by type I's. Prices biases were apparently not due to simple one or two type I's buying units, because about 80 percent of the type I subjects held any units, comparéd to roughly 30 percent of the type II subjects. Significance tests using mean data from each experiment strongly reject the Bayesian theory against the alternative of exact representativeness.[7] Such cross-experiment tests are especially reliable because we can be confident that different experiments are genuinely independent because they contain different subjects.

The smaller amount of data from 2-red periods (the bottom panel of Table 4) are not very conclusive. The Bayesian theory predicts type I's will hold, exact representativeness predicts type II's, and holdings are about equal. This corroborates the finding from price data that exact representativeness has little effect in 2-red periods.

The results of Duh and Sunder (1986) are worth summarizing at this point. In their experiments, the two states (called R and W) are two bingo cages containing 16 red and 4 black balls (R) and 4 red and 16 black balls (W). The prior $P(R)$ varied from .65 to .85 across experiments, since their main concern was whether subjects ignored prior probabilities. One ball is drawn from whichever cage (state) is chosen (so there is no possibility of exact representativeness). They find that when an R is drawn, prices are close to Bayesian. When a W was drawn, the Bayesian theory predicted about as well as a base-rate ignorance theory (denoted $NBR2$) in which $P(R)$ and $P(W)$ are judged to be equal, and an extreme version of representativeness in which $P(W)$ is judged to be

[7]We can test the hypothesis that the average percentage holding of type I's was 50 percent by assuming the fractions across experiments 1 to 5, $9r$ and $10h$ are normally distributed (the t-statistic is 9.28). The more conservative binomial test of successes yields a significance level less than 1 percent. For experienced subjects these statistics are 10.33 and 6 percent.

TABLE 4—HOLDINGS OF UNITS AT PERIOD END, BY TRADER TYPE

Experiment (n = no. of periods)	Type I		Type II	
	1-Red Periods			
Theories predicting Each Type to Hold:	Bayesian, Exact Representativeness, Overreaction		Bayesian, Conservatism, Base-Rate Ignorance	
Inexperienced Subjects	Fraction Holding Any	Fraction Held	Fraction Holding Any	Fraction Held
1 ($n=5$)	.76	.85	.24	.15
2 ($n=7$)	.76	.82	.50	.18
3 ($n=9$)	.94	.75	.42	.25
4 ($n=12$)	.67	.73	.38	.27
5 ($n=10$)	.76	.64	.56	.36
$9r$ ($n=8$)	.91	.85	.30	.15
$10h$ ($n=7$)	.69	.85	.37	.15
Means All Periods:	.787	.784	.396	.216
Early Periods:	.82	.73	.53	.27
Late Periods:	.76	.82	.31	.18
Experienced Subjects				
$11x$	.91	.93	.25	.07
$13x$	.95	.78	.56	.22
$14x$	.50	.95	.10	.05
$15xh$	.91	.97	.09	.03
Means All Periods:	.818	.908	.250	.092
Early Periods:	.88	.87	.29	.13
Late Periods:	.76	.94	.21	.06
	2-Red Periods			
Theories Predicting Each Type to Hold	Bayesian Conservatism		Exact Representativeness, Overreaction, Base-Rate Ignorance	
6 ($n=11$)	.53	.36	.75	.64
7 ($n=11$)	.57	.60	.45	.40
8 ($n=8$)	.71	.49	.83	.51
$12x$ ($n=13$)	.76	.40	.88	.60
Means All Periods:	.643	.463	.728	.538
Early Periods:	.64	.42	.77	.58
Late Periods:	.66	.52	.67	.48

one (denoted *NBR*1). They do not estimate the degree of price bias parametrically, but it seems to be smaller in magnitude than the biases observed here. They conclude, "Although the Bayesian model performs best among the four models in its ability to predict transaction prices, the observed market behavior still deviates from the Bayesian prescription." I suspect the Bayesian model predicts better in their experiments than in mine because the exact representativeness in my experiments is a stronger psychological force than the base-rate ignorance in theirs.

C. *Further Controls for Risk and Incentives*

The analyses of prices and allocations lean heavily on the assumption that traders are risk neutral, so that they trade at expected values. If traders are risk seeking, prices will be above expected values. The higher prices observed in 1-red periods could therefore reflect risk seeking by Bayesian traders rather than judgment bias by risk-neutral traders.

This explanation is unlikely for several reasons. First, the Arrow-Pratt risk pre-

mium, which measures the approximate degree to which prices depart from expected values because of risk seeking, depends only on the variance of an asset's value (and possibly its mean) and the shape of traders' utility functions. The mean and variance of the value of units are identical for type I and type II traders in 1-red periods, so their risk premia should be equal (assuming no systematic differences in utility functions). Therefore, the Bayesian prediction that type I and type II traders hold equal amounts of units should be true even if traders are not risk neutral; but the equal holdings prediction is strongly rejected.

Second, most attempts at measuring risk tastes in experimental settings find evidence of risk aversion rather than risk seeking (for example, James Cox, Smith, and James Walker, 1985; Smith, Gerry Suchanek, and Williams, 1987). Third, the allocation data show that about 80 percent of the type I traders are holding units at the high prices in 1-red periods. It seems unlikely that almost every type I trader in every experiment would be risk seeking. Fourth, the data from all four samples can be used to estimate the degree of risk seeking implicit in prices, assuming a specific utility function. Adjusting the apparent price biases in 1-red periods for risk does reduce them by about two-thirds, but not quite to zero.[8]

More direct evidence of whether risk seeking can explain the biases comes from a control experiment (denoted $9r$) in which risk neutrality was induced by design (see Alvin Roth, 1983; Joyce Berg et al., 1986; though, see Cox et al., 1985). Traders accumulated earnings in francs but the francs were not converted into dollars at the end of the experiment. Instead, traders were paid \$15 plus a \$50 bonus if a uniformly distributed five-digit number between 0 and 50,000 was *below* their amount of earnings. Each franc they earned then raised their probability of winning the \$50 prize by 1/50,000; so francs were like units of probability. Since assets are lotteries over possible amounts of francs, and francs are probability units, assets are like compound lotteries. If traders satisfy the reduction of compound lotteries axiom in expected utility theory, they should regard a gamble with an expected franc value of G as identical to a certain payment of G francs, so they should act as if they are risk neutral toward francs. If biases observed in earlier experiments were due to risk seeking, those biases should disappear in experiment $9r$.

A second control experiment (denoted $10h$) used a "high-stakes" dollar-per-franc conversion rate of \$.005 rather than \$.001. Subjects in this experiment made about \$20 per *hour*. Experiment $15xh$ used the same level of high stakes with experienced subjects. If apparent price biases are due to insufficient incentive to think carefully about probabilities, biases should be smaller in experiments $10h$ and $15xh$.

Figure 4 shows the mean prices from the risk-control experiment $9r$ (thick line) and the high-stakes experiment $10h$ (thin line).[9] Compared to prices from inexperienced subjects shown in Figure 2, prices in these experiments are extremely close to the Bayesian expected values, except in 1-red periods. Price regression results and allocations (in Tables 3 and 4) suggest the exact representativeness bias in 1-red periods is highly significant. Therefore, biases in 1-red periods

[8]The value of the risk-seeking constant A was estimated in each experiment, assuming both constant absolute (CARS) and constant relative risk seeking (CRRS). The value of A was chosen to minimize the absolute deviations between observed price bias from Table 3 and the bias predicted by the Arrow-Pratt risk premium with parameter A, summed across the four possible samples. Weighting deviations by the number of trades in each sample minimized risk-adjusted biases better than not weighting them. The CARS and CRRS models fit almost identically. Using CARS, risk-adjusted biases in 1- and 2-red periods averaged 13.8 and -15.6 francs (experiments 1 to 8) and 4.5 and -6 francs (experiments $11x$ to $15xh$). In experiments $9r$ and $10h$ risk adjustment actually increased 1-red biases to 55.5 and 56.4 francs. Furthermore, in experiment $9r$, the estimated A was about equal in magnitude to A's in other experiments, though it should be zero in theory.

[9]The lines end abruptly because each experiment has a different number of periods of each sample. There are five 0-red periods in $10h$, for instance, and only three in $9r$. Also, the spike in the second 1-red period of experiment $9r$ was a short burst of irrational buying at very high prices, which defies explanation.

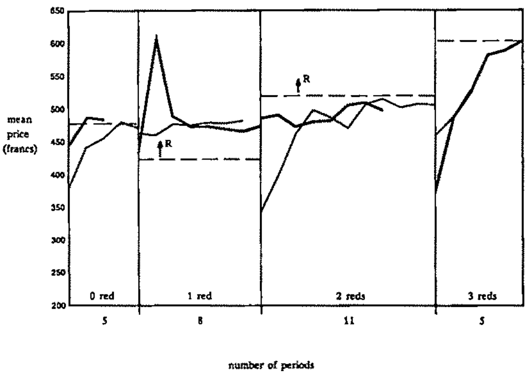


FIGURE 4

in other experiments are probably not due to risk seeking or insufficient motivation. At the same time, the control experiments give evidence *against* the exact representativeness prediction in 2-red periods.

D. *Individual Judgments and Market Prices*

The point of experiments like these is to compare behavior of individuals with behavior of markets in which individuals participate. So far there has been only an assumption individuals will err in using Bayes' rule, but no direct comparison between individuals and the market. However, we can make such a comparison because subjects did make individual probability judgments before trading began (except in experiments 1 and 2).

Judgments were rewarded with a quadratic scoring rule, with money incentives for accuracy.[10] The scoring rule is incentive compatible assuming risk neutrality (subjects should report their true subjective probabilities), but nonrisk neutrality will cause judgments to deviate from true beliefs. Subjects were given 10 to 20 three-ball samples from the bingo cages, with instant feedback about whether X or Y occurred. After completing the scoring-rule exercise, subjects were informed that they would ranked according to their earnings from the scoring-rule exercise, from 1 to N. They were told to predict their rank, choosing exactly one number between 1 and N, and they were paid \$5 if their rank was exactly correct.

We can compare the average scoring-rule judgment with a probability estimate imputed from the equilibrium price bias. For instance, in experiment 3 the estimated bias in 0-red periods was -11.09 francs (see Table 3). Since type I traders were holding in these periods, and their payoffs range from 200 ($P(X)=0$) to 500 ($P(X)=1$), the probability scale naturally corresponds to a 300-franc price scale from 200 to 500. A bias of -11.09 francs implies a probability judgment of $P(X/0\text{-red})$ that is $-11.09/300$, or $-.037$, different from the Bayesian posterior of .923. Probabilities were imputed from market prices for each sample and each experiment, using the estimated biases from Table 3.

Average individual probabilities from the scoring rule and probabilities imputed from market prices, averaged across experiments, are shown in Table 5. Both kinds of probabilities are close to Bayesian in 0- and 3-red samples. In 1- and 2-red samples, the individuals' probability estimates are closer to Bayesian than the market prices are,[11] but the gap is smaller with experienced subjects.

It seems that for exactly representative samples, markets are often *more* biased than

[10] Samples of three balls were drawn, exactly as in determining states, and subjects were asked to choose a two-digit "decision number" from 00 to 99. Define that number, divided by 100, as D. If event X occurred, subjects were paid $2D-D^2$ dollars. If event Y occurred, subjects were paid $1-D^2$ dollars. Subjects were shown a table of the possible numerical payoffs. If a subject's true subjective probability of X occurring was S, and she choose D, her expected payoff was $S(2D-D^2)+(1-S)(1-D^2)$. This payoff has a maximum at $D^*=S$, that is, subjects should truthfully choose their subjective probabilities as their decision numbers, except for risk aversion. If subjects are risk averse (risk seeking), their reported probabilities will be biased toward (away from) .5.

[11] The differences between averaged scoring-rule judgments and probabilities implicit in market prices are highly significant by parametric t-tests, or by nonparametric matched-pairs or rank-sum tests, except in 2-red periods with inexperienced subjects.

TABLE 5—AVERAGE PROBABILITY JUDGMENTS OF INDIVIDUALS AND PROBABILITIES IMPLICIT IN MARKET PRICES (EXPRESSED AS DEVIATIONS FROM THE BAYESIAN POSTERIOR $P(X/\text{SAMPLE})$)

Sample	0-Red	1-Red	2-Red	3-Red
Direction of Deviation Predicted by Exact Representativeness:	0	+	−	0
Inexperienced Subjects (8 Experiments)				
Individual Mean	−.009	−.030	−.031	−.022
(Standard Deviation)	(.044)	(.087)	(.033)	(.076)
Market Prices Mean	−.008	+.141	−.100	−.035
(Standard Deviation)	(.062)	(.081)	(.193)	(.073)
Experienced Subjects (5 Experiments)				
Individual Mean	+.037	−.026	−.043	−.084
(Standard Deviation)	(.022)	(.052)	(.061)	(.069)
Market Prices Mean	+.007	+.059	−.081	−.016
(Standard Deviation)	(.092)	(.049)	(.044)	(.117)

individuals are. One explanation is that market prices are determined by one or two highly biased traders, but almost all traders were holding units at the biased prices. Another possibility is that the market mechanism and the quadratic scoring rule simply elicit different probability judgments.

One consolation is that the biases shrink with experience. A closer look at individual data may suggest why. For market prices to be less biased than individuals, traders who are less biased must exert more influence on the market price. There is no external market to evaluate whether traders are unbiased and allocate more trading capital to them. Therefore, to exert more influence the traders who are less biased must realize they are less biased, and trade more aggressively.

Whether traders realize their relative ability at probability judgment can be measured by whether their predicted ranks in the scoring-rule exercise are correlated with their actual ranks. The two sets of ranks were somewhat correlated—averaging .49 for inexperienced subjects and .30 for experienced subjects[12]—so subjects do have some self-insight. However, predictions about relative ability are not highly correlated with the amount of arbitrage (defined as buying and selling in the same period). Those correlations averaged −.09 for inexperienced subjects, and .23 for experienced subjects. Furthermore, actual ranks and arbitrage were uncorrelated (.05 and −.12) with both inexperienced and experienced subjects. It seems that aggressive trading, as measured by arbitrage, is not something inexperienced subjects do only because they think they are better probability judges than others.

III. Conclusion and Future Research

In many experiments subjects do not follow the laws of probability, particularly Bayes' rule. However, subjects in these experiments are often unpaid and given little practice making judgments. In markets, traders often have incentives and experience, and people who are good at estimating probabilities can often exert more force on prices. Therefore, biases in *individual* judgments need not affect prices and allocations in *markets*.

Whether biases affect market outcomes is tested in a series of simple experimental markets. In the markets, traders exchange units of an asset that pays a state-dependent dividend. A random device yields sample evidence about which state has occurred. Traders' demand for assets depends upon

[12]These are high correlations considering that the range of the predicted rank variable was restricted by subjects' optimism about their ranks. Sixty-two of 74 inexperienced subjects (84 percent) thought they were in the top 50 percent in scoring-rule earnings, compared to 23 of 40 experienced subjects (58 percent). Apparently optimism is nearly erased after one experiment.

their judgments about posterior state probability. If the market functions as if traders are Bayesians, a certain pattern of prices and allocations is predicted to occur. But if traders overestimate P(state/sample), relative to the Bayesian posterior, when the sample exactly matches the contents of a bingo cage that represents the state, then different prices and allocations will occur. This competing theory is called "exact representativeness." It is less useful than the Bayesian theory because it does not predict prices when samples do not exactly match states, but it does have some bite. Other non-Bayesian psychological theories can be defined too.

In eight experiments with inexperienced subjects, prices tend toward the Bayesian predictions, but there is some evidence of exact representativeness bias in prices and allocations. However, the degree of bias is small, and it is even smaller in experiments with experienced subjects. All other non-Bayesian theories can be rejected.[13] Furthermore, the Bayesian theory predicts prices remarkably well when the exact representativeness theory does not apply.

In most experiments, biases are statistically significant for only one of the two samples (the 1-red sample) in which exact representativeness predicts bias. Indeed, if the reader values the only experiment with controls for risk seeking (9*r*), exact representativeness predicts no better than chance: it predicts the significant bias in the 1-red period correctly, but it predicts the wrong sign on the significant bias in the 2-red period.

It is easy to imagine other market settings in which unbiased traders could correct market biases completely.[14] Some of these settings are the subject of ongoing research. However, if one pretends to not know the results, it is easy to imagine that biases could have been entirely eliminated in these experiments, too.

Whether the exact representativeness biases in 1-red periods are significant depends upon your yardstick of significance. By one overworked yardstick, the statistical test of whether they could be due to chance, the biases in 1-red periods are highly significant. The possibility of excess profits is an important yardstick in economics. There are apparently loss of profits to be earned from exploiting biased subjects, since they overpay by roughly $.20 per trade (a few dollars per experiment) in 1-red periods of the high-stakes experiments 10*h* and 15*xh*. Excess profits are a lot smaller, only about $.03 per trade, in the other experiments. On the probability yardstick the biases are errors of about .10, which are large if your purpose is testing students' ability to make exact Bayesian calculations and small if your purpose is comparing these biases with errors found in other studies.[15]

Of course, if the stakes were large enough or (perhaps more importantly) traders had enough experience, the apparent biases might disappear entirely. Therefore, we should hesitate to generalize these results to the New York Stock Exchange (though some have tried[16]), but the results may generalize to settings in which stakes are relatively small and agents have little experience in a repeated situation. For instance, consumers might judge the quality of a new product by how much the product's packaging or advertising resembles that of well-known products. Financial journalists sometimes argue a depression is ahead because a pattern of economic indicators resembles a pattern from

[13] If subjects tended to ignore or underweight the unequal prior probabilities of the states, then 2-red biases would be larger than 1-red biases. Exactly the opposite is true. Notice also that overreaction predicts reasonably well in 1- and 2-red samples, when it overlaps with exact representativeness, but it predicts poorly in 0- and 3-red samples.

[14] For instance, if biases caused prices to be lower than expected values, then unbiased traders would pay higher prices than biased traders, effectively setting the market price, so prices might appear unbiased.

[15] For instance, in the well-known blue-green taxi problem (for example, Tversky and Kahneman, 1982b), the Bayesian posterior is around .4 but subjects often answer .80 because they ignore the low base rate of one type of taxi.

[16] Recall Arrow's (1982) suggestion cited above. Werner DeBondt and Richard Thaler (1985) also found empirical support for the representativeness prediction that investors do not expect regression in extreme earnings announcements.

before the Great Depression. (Whether such opinions affect market behavior is debatable.) The belief that the future is likely to be representative of the past could cause a failure to anticipate regression effects (Tversky and Kahneman, 1982b): Forgetting about regression, consumers may avoid all Hyatt hotels or DC-10's after an accident involving one of them; or studios might make movie sequels that are consistently unprofitable. The winner's curse in common-value auctions (see John Kagel and Dan Levin, 1986) might be caused by a heuristiclike representativeness. These conjectures, whether plausible or not, illustrate how representativeness bias akin to that observed in the experiments could affect economic outcomes in natural settings.

There are several directions for future experiments. Institutional extensions of these markets, like short selling or a parallel market for information about probabilities, might eliminate biases entirely. Experiments in which other judgment biases could affect markets might be interesting too (for example, myself, George Loewenstein, and Martin Weber, 1987). A program of empirical work, including both experiments and extending experimental results to natural settings, could establish what kinds of irrationality seem to persist under the incentives and learning opportunities present in natural markets. Such data might lead to economic theory that uses evidence of systematic irrationality to make better predictions.

REFERENCES

Arkes, Hal R. and Hammond, Kenneth R., *Judgment and Decision Making: An Interdisciplinary Reader*, Cambridge: Cambridge University Press, 1986.

Arrow, Kenneth, "Risk Perception in Psychology and Economics," *Economic Inquiry*, January 1982, *20*, 1–9.

Berg, Joyce E., Daley, Lane A., Dickhaut, John W. and O'Brien, John R., "Controlling Preferences for Lotteries on Units of Experimental Exchange," *Quarterly Journal of Economics*, May 1986, *101*, 281–306.

Blaug, Mark, *The Methodology of Economics: or How Economics Explain*, Cambridge: Cambridge University Press, 1980.

Camerer, Colin, Loewenstein, George and Weber, Martin, "The Curse of Knowledge in Economic Settings: An Experimental Analysis," Wharton Risk and Decision Processes Center, Working Paper No. 87-09-07, 1987.

Cox, James, Smith, Vernon and Walker, James, "Experimental Development of Sealed-bid Auction Theory: Calibrating Controls for Risk Aversion," *American Economic Review*, May 1985, *75*, 160–65.

DeBondt, Werner F. M. and Thaler, Richard, "Does the Stock Market Overreact?," *Journal of Finance*, July 1985, *40*, 793–805.

Duh, Rong Ruey and Sunder, Shyam, "Incentives, Learning and Processing of Information in a Market Environment: An Examination of the Base-Rate Fallacy," in S. Moriarty, ed., *Laboratory Market Research*, Norman, OK: Center for Economic and Management Research, University of Oklahoma, 1986.

Easley, David and Ledyard, John, "Theories of Price Formation and Exchange in Double Oral Auctions," Social Science Working Paper No. 611, California Institute of Technology, 1986.

Edwards, Ward, "Conservatism in Human Information Processing," in B. Kleinmuntz, ed., *Formal Representation of Human Judgment*, New York: Wiley & Sons, 1968.

Friedman, Daniel, "On the Efficiency of Experimental Double Auction Markets," *American Economic Review*, March 1984, *74*, 60–72.

Friedman, Milton, *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Grether, David M., "Bayes' Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics*, November 1980, *95*, 537–57.

Hey, John D., "Search for Rules for Search," *Journal of Economic Behavior and Organization*, March 1982, *3*, 65–81.

______, "Still Searching," *Journal of Economic Behavior and Organization*, March 1987, *8*, 137–44.

Kagel, John H. and Levin, Dan, "The Winner's Curse and Public Information in Common Value Auctions," *American Economic Review*, December 1986, *76*, 894–920.

Kahneman, Daniel, Slovic, Paul and Tversky, Amos, *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, 1982a.

Kmenta, Jan, *Elements of Econometrics*, New York: Macmillan, 1971.

Plott, Charles R. and Wilde, Louis L., "Professional Diagnosis vs. Self-diagnosis: An Experimental Examination of Some Special Features of Markets with Uncertainty," in V. Smith, ed., *Research in Experimental Economics*, Vol. 2, Greenwich: JAI Press, 1982.

Roth, Alvin, "Toward a Theory of Bargaining: An Experimental Study in Economics," *Science*, May 13, 1983, *220*, 687–91.

Russell, Thomas and Thaler, Richard, "The Relevance of Quasi Rationality in Competitive Markets," *American Economic Review*, December 1985, *75*, 1071–82.

Slovic, Paul, Fischhoff, Baruch and Lichtenstein, Sarah, "Cognitive Processes and Societal Risk Taking," in J. S. Carroll and J. W. Payne, eds., *Cognition and Social Behavior*, Hillsdale, NJ: Erlbaum, 1976.

Smith, Vernon L., "Microeconomic Systems as an Experimental Science," *American Economic Review*, December 1982, *72*, 923–55.

______ **, Suchanek, Gerry and Williams, Arlington,** "Bubbles, Crashes, and Endogeneous Expectations in Experimental Asset Markets," Department of Economics Working Paper No. 86-2, University of Arizona, 1987.

______ **and Williams, Arlington,** "The Boundaries of Competitive Price Theory: Convergence, Expectations, and Transaction Costs," in L. Green and J. Kagel, eds., *Advances in Behavioral Economics*, Vol. 2, Norwood, NJ: Ablex Publishing, in press.

Tversky, Amos and Kahneman, Daniel, (1982a) "Judgments of and by Representativeness," in D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

______ **and** ______, (1982b), "The Evidential Impact of Base Rates," in D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Wilson, Robert, "Incentive Efficiency of Double Auctions," *Econometrica*, September 1985, *53*, 1101–115.

The Cost of Regulation: OSHA, EPA and the Productivity Slowdown

By WAYNE B. GRAY*

The slowdown in productivity growth in the U.S. economy during the 1970's has been a matter of great concern to policymakers, associated as it is with inflation, unemployment, and declining real wage growth. This paper examines the impact on productivity growth of government regulation, specifically worker health and safety regulation by the Occupational Safety and Health Administration (OSHA) and environmental regulation by the Environmental Protection Agency (EPA). Looking at data for 450 manufacturing industries between 1958 and 1978, the study finds a large, negative relationship between such regulation and productivity growth. Using these results, about 30 percent of the decline in productivity growth in manufacturing during the 1970's may be attributed to such regulation.

Several previous studies have looked at the contribution of regulation to the productivity slowdown. Many of these have inferred that the contribution must be small, on the basis of the relatively small amount spent on complying with such regulations. Edward Denison (1979) estimates that only about 16 percent of the productivity slowdown in the 1972–75 period was due to regulation (.35 percentage points out of a slowdown of 2.17 percentage points). Paul Portney (1981) notes that little of GNP is spent on pollution control (under 2 percent), concluding that therefore pollution regulations could have little effect on productivity growth. Norsworthy et al. (1979) also find a small impact of pollution-abatement capital expenditures on productivity growth.

Studies based on econometric estimation of the regulation-productivity relationship have found a wide range of results. Gregory Christainsen and Robert Haveman (1981) find regulation reduced labor productivity growth by .27 percentage points, using time-series data and measures of total federal regulation. Robert Crandall (1981) finds a strong relationship between pollution-abatement capital and productivity growth, but this relationship disappears when a measure of energy intensity is included. Robin Siegel (1979) observes a significant contribution (.5 percentage points) from pollution control expenditures to the productivity slowdown for 1965–73, but not for later years. Finally, Frank Gollop and Mark Roberts (1983) examine data for a set of electric utilities and find that regulation of emissions had a large impact on total factor productivity growth, lowering it for regulated firms by .59 percentage points.

Many other factors might help to explain the productivity slowdown, including the rise in energy prices, the long and severe recession, and declines in research and development expenditures. There have been a variety of studies examining the contributions of each factor to the slowdown. They generally conclude that many factors contributed to the slowdown, but that a sizable fraction of the slowdown remains unexplained by the estimated contributions of all the factors considered.[1]

*Department of Economics, Clark University, Worcester, MA 01610. This research was partially supported by the Sloan Foundation and the National Bureau of Economic Research. I thank seminar participants at the National Bureau and anonymous referees for helpful comments. I also thank various individuals in OSHA, EPA, and the Commerce Department for providing data.

[1]Exceptions to this tendency are not uncommon, with several studies (such as Michael Darby, 1984, and Thomas Weisskopf et al., 1983) finding that the particular factor under consideration explains nearly all (or even more than 100 percent of) the productivity slowdown.

I. The Model

This paper concentrates on total factor productivity (TFP) measures of productivity growth, which consider the contribution of all productive inputs to output growth. Given a simple production function,

$$(1) \qquad Y = T * F(X_1, \ldots, X_N),$$

where output Y depends on the level of productivity T (assumed to be Hicks-neutral) and inputs X_i, we can calculate TFP growth (τ) as

$$(2) \qquad \tau = dy - \Sigma \alpha_i \, dx_i,$$

where α_i is the share of input i in total cost, and dy and dx_i are the growth rates of Y and X_i. This method of calculating productivity growth, known as growth accounting, requires no estimation of the production function, but cannot test for changes in the productivity of different inputs over time.

Regulation could affect measured productivity growth by requiring firms to use some inputs for compliance. If a firm uses R_i of each input to comply with regulations, but TFP is calculated without recognizing this, X_i, α_i, and τ will be mismeasured as X_i', α_i' and τ':

$$(3) \qquad X_i' = X_i + R_i, \quad \alpha_i' = p_i X_i' / p_y Y,$$

$$\text{and} \qquad \tau' = dy - \Sigma \alpha_i' \, dx_i'.$$

If the fraction of each input used in compliance is $\theta_i (= R_i / X_i')$ we have the following relationship between measured (τ') and true (τ) TFP growth:

$$(4) \qquad \tau' - \tau = \Sigma \alpha_i x_i - \Sigma \alpha_i' x_i'$$
$$\cong \Sigma \alpha_i' (x_i - x_i') \cong -\Sigma \alpha_i' \theta_i,$$

where the quality of the approximation depends on θ_i being close to 0.[2] This "measurement effect" (so called because the inputs actually contributing to output are being mismeasured) leads a growth accounting measure to understate true productivity growth. We can further simplify this:

$$(5) \qquad \Sigma \alpha_i' \theta_i = \Sigma (p_i X_i' / p_y Y)(R_i / X_i')$$
$$= (\Sigma p_i R_i) / p_y Y = \theta,$$

showing that it is not necessary to know how much of each input is used for compliance; the share of compliance costs in total cost (θ) is enough.

In addition to the "measurement effect," regulation might have a "real effect" on productivity. It could impose constraints on the firm's choice of production processes, make it harder to take advantage of new innovations, cause firms to lower new investment by increasing uncertainty, or otherwise reduce the productivity of other (noncompliance) inputs.[3] If we measure productivity growth without (0) and with (1) regulation imposed, we get

$$(6) \qquad d\tau' = \tau_1' - \tau_0' = \tau_1 - \tau_0 - \theta,$$

where $\tau_1 - \tau_0$ is the real effect and θ is the measurement effect.

If we observe $d\tau'$ and θ for many different firms, indexed by j, we could estimate an equation like the following:

$$(7) \qquad d\tau_j' = \mu - \beta(\theta_j) + \varepsilon_j.$$

Here μ and ε_j allow for influences other than regulation to affect productivity growth (economywide and firm-specific, respectively). If regulation had only a measurement effect, the regression suggested by (7) would yield $\beta = 1$. If regulation does have a real effect, we would get $\beta > 1$, as firms facing greater regulation would have lower true productivity growth. If a measure of compliance costs were not available for a particular

[2] This ensures the α_i' is close to α_i and that $x_i - x_i'$ is close to $-\theta_i$.

[3] Nicholas Ashford and George Heaton (1983) and Fred Hoerger et al. (1983) examine these issues for the chemical industry. Kip Viscusi (1983) develops a model of how regulation could affect investment.

type of regulation, some alternative measure of regulation could be used (replacing θ_j), but the effect on productivity could not be separated into real and measurement components.

All of the foregoing discussion has been in terms of individual firms' productivity growth. Unfortunately, firm-specific information linking productivity growth with measures of regulation is not available. Instead, I use industry-level data on output and inputs in a growth accounting calculation of TFP growth for each industry. This is then related to measures of the amount of regulation faced by each industry, as in equation (7), with controls for other factors that might affect industry productivity growth.

Some previous studies have estimated industry production or cost functions to look at productivity growth, rather than simply calculating TFP. An example of this is found in Dale Jorgenson (1984), which looks at the relation between energy usage and productivity growth. This allows the investigation of the effect of regulation on the productivity of different inputs, but is not done here for a number of reasons. First, the amount of regulation faced by an industry is not determined by the industry, but by the regulatory agency, so it can properly be treated as exogenous to the industry (unlike energy usage). Also, a proper model of the response by an industry to regulation would be quite complicated, depending on past and expected future costs of compliance, penalties for noncompliance, and enforcement efforts (much of which cannot be measured here). Finally, the goal here is to examine how much of the productivity slowdown might be attributed to regulation, so the first-order relation between regulation and productivity seems to be the appropriate level of analysis.

II. Data Description

The data set I created for this analysis is the first to combine extensive information on both productivity growth and regulation. The data cover the entire U.S. manufacturing sector, divided into 450 separate industries. Annual productivity growth for each industry is calculated from 1958 to 1978, based on a growth accounting model with five inputs (see the Data Appendix for data sources). TFP growth is calculated as real output growth minus real input growth (the real growth of each input, weighted by its cost share):

$$(8)\quad \tau_t = (\log Y_t - \log Y_{t-1}) - \Sigma\left[(\alpha_{it} + \alpha_{it-1})/2 \times (\log X_{it} - \log X_{it-1})\right].$$

The measure of EPA regulation of each industry is the industry's annual operating cost associated with pollution control. The data are based on the Pollution Abatement Costs and Expenditures survey of about 20,000 establishments, taken annually by the Bureau of the Census since 1973. Following equation (5), the compliance costs are divided by the value of shipments for the industry, yielding θ_j.

There is no usable data on costs to individual industries of complying with OSHA regulation.[4] Instead, a measure of the enforcement effort directed by OSHA toward each industry is used. OSHA's Management Information System, which tracks OSHA inspections for agency review purposes, identifies the industry and number of employees for each inspected establishment. The number of workers inspected is aggregated each year by industry and divided by total industry employment, yielding the fraction of industry employment in plants inspected by OSHA that year.

The use of an enforcement measure does not allow us to separate OSHA's effect on productivity into real and measurement components. However, enforcement effort is likely to be positively correlated across in-

[4] There is an annual McGraw-Hill survey on capital spending that asks what fraction of total capital spending is allocated to worker safety and health. Unfortunately, it has little industry detail, covers only a few hundred firms, and shows little correlation with either productivity growth or other measures of OSHA regulation.

dustries with compliance cost (establishments in industries without serious health and safety risks have no need to expend resources improving performance, and are less likely to be inspected). Also, enforcement may itself impose costs on establishments (for example, OSHA inspections disrupting the normal production routine). Thus differences in enforcement effort across industries should measure differences in the impact of OSHA regulation on those industries.

III. Results

The basic variables used in the analysis are presented in Table 1. To reduce the impact of strong cyclical fluctuations in productivity, average TFP growth is calculated for periods covering several years, chosen to match the cycle of productivity fluctuations from peak to peak. The measure of the productivity slowdown for each industry is the change in average annual TFP growth between the 1959–69 period and the 1973–78 period (*TFPCHG*).[5] The earlier period was chosen to end before the regulatory agencies studied here began operating, to ensure that the measures of levels of regulation in the later period would also measure changes in regulation from the earlier period.[6] The level of productivity growth in the later period (*TFP7378*) is also examined, to see whether results for the TFP slowdown are due to faster TFP growth in the earlier period or slower TFP growth in the later period.

TABLE 1—DESCRIPTIVE STATISTICS
(Full Sample, 450 industries)

Variable	Mean	Description
TFPCHG	−.0146 (.032)	Change in annual TFP growth rate: 1959–69 to 1973–78
TFP7378	−.0054 (.029)	Annual TFP growth rate: 1973–78
OSHINS	.5404 (.665)	Average OSHA employee inspection rate: 1974–78
PAOC	.0029 (.005)	Average pollution-abatement operating costs as share of total cost: 1974–78
SHEN	.0164 (.022)	Average cost share of energy: 1969–73
SHCAP	.2630 (.078)	Average cost share of capital: 1969–73
GPLCHG	−.0216 (.053)	Change in growth rate of production worker hours: 1959–69 to 1973–78
TFPCHGX	.0032 (.035)	Change in annual TFP growth rate: 1959–63 to 1963–69

Note: Standard deviations are shown in parentheses.

The explanatory variables measure both regulation and other factors that might affect productivity growth.[7] The regulation measures, *OSHINS* and *PAOC*, are averaged over five years of data in the later period. Two input cost shares, *SHEN* and *SHCAP*, are used to test the possibility that industries which are energy- or capital-intensive suffered greater productivity slowdowns than average. Industries that experienced faster employment growth in the 1970's than in the 1960's (measured by *GPLCHG*) might have experienced less of a productivity slowdown. Finally, industries with declining productivity growth in the 1960's (measured by *TFPCHGX*) might have continued declining in the 1970's.

[5] The results are not materially affected by extending the later time period to 1980 or changing the earlier period, but are somewhat sensitive to the choice of 1973 as the starting year.

[6] There certainly was some such regulation before OSHA and EPA, but most of it was administered by state agencies that (with few exceptions) had little enforcement power. As long as the regulation in the earlier and later periods is similarly distributed across industries, it will not affect the estimated impact of regulation. To see this, suppose that each industry faced half as much regulation in the earlier period as it did in the later one. The regulation measure (which purports to measure the change in regulation) will be twice as big as it should be, but its coefficient will therefore be cut in half, exactly canceling out.

[7] Several other regulation measures were tested, but not presented, as they did not affect the basic results. They included capital expenditures on pollution abatement (taken from the same PACE survey that was the source for the *PAOC* variable) and EPA inspections of establishments for violation of air-quality standards (based on information from EPA's Compliance Data System, similar to the OSHA data). Both were negatively correlated with productivity growth (in levels and changes), but are not included because they are less comprehensive than *PAOC*.

TABLE 2—CORRELATIONS

	TFPCHG	*TFP7378*	*OSHINS*	*PAOC*	*SHEN*	*SHCAP*	*GPLCHG*	*TFPCHGX*
TFPCHG	1.0	.86	−.14	−.17	−.15	−.15	.15	.17
TFP7378		1.0	−.16	−.20	−.17	−.09	.09	.14
OSHINS			1.0	.33	.18	−.08	.01	−.09
PAOC				1.0	.66	.09	.03	−.03
SHEN					1.0	.18	.07	.02
SHCAP						1.0	.11	.03
GPLCHG							1.0	.08
TFPCHGX								1.0

TABLE 3—INITIAL REGRESSION RESULTS (Full sample, $n = 450$)

Model	Constant	*OSHINS*	*PAOC*	*SHEN*	*SHCAP*	*GPLCHG*	*TFPCHGX*	$R^2(SSE)$
Dependent Variable = *TFPCHG*								
A1	−.0110	−.0068	–	–	–	–	–	.020
	(.0019)	(.0018)						(.445)
A2	−.0113	–	−1.17	–	–	–	–	.029
	(.0017)		(.40)					(.441)
A3	−.0094	−.0046	−0.95	–	–	–	–	.037
	(.0020)	(.0017)	(.42)					(.437)
A4	.0100	−.0050	−0.60	−.078	−.066	.100	.135	.111
	(.0053)	(.0015)	(.55)	(.093)	(.019)	(.028)	(.049)	(.403)
Dependent Variable = *TFP7378*								
B1	−.0015	−.0072	–	–	–	–	–	.027
	(.0018)	(.0017)						(.374)
B2	−.0035	–	−1.28	–	–	–	–	.041
	(.0015)		(.30)					(.368)
B3	.0002	−.0048	−1.05	–	–	–	–	.052
	(.0018)	(.0015)	(.31)					(.364)
B4	.0102	−.0049	−0.69	−.100	−.032	.055	.101	.086
	(.0050)	(.0015)	(.45)	(.080)	(.018)	(.030)	(.048)	(.351)

Contribution of Regulation to TFP Slowdown between 1958–69 and 1973–78

	Contribution to Slowdown[a]			Fraction of Slowdown[b]		
	OSHINS	*PAOC*	*Total*	*OSHINS*	*PAOC*	*Total*
A1	−.36	–	−.36	.25	–	.25
A2	–	−.34	−.34	–	.23	.23
A3	−.25	−.28	−.53	.17	.19	.36
A4	−.27	−.17	−.44	.19	.12	.31

Note: Standard errors are shown in parentheses.
[a] The predicted impact of the mean value of each regulation measure (.54 for *OSHINS* and .0029 for *PAOC*) on *TFPCHG*, measured in percentage points per year.
[b] Calculated as (contribution to slowdown)/(mean slowdown = −1.46).

Several things can be learned from the correlations in Table 2. The regulation measures are negatively correlated with the productivity measures (as expected). They are positively correlated with each other, so analyzing only one measure would overstate its effect on productivity. Finally, they are correlated with some of the other factors that might affect productivity growth, so that omission of these factors could bias the estimated effect of regulation on productivity. Of course, any remaining factors (regulatory or otherwise) that have not been included here could also bias the results.

The regression results found in Table 3 show the connection between the regulation

measures and the productivity slowdown.[8] When analyzed alone, each regulation measure has a significant negative coefficient. When both regulation measures are included, their coefficients fall (as expected from the correlations earlier), but generally remain significant. With all of the other factors included, the effect of *PAOC* falls by about 40 percent (and is no longer significant), but *OSHINS* remains strong.

One feature of these results is the similarity between the coefficients on the regulation measures in the regressions explaining the change in the rate of productivity growth (A) and those explaining the rate of productivity growth (B). This means that the regulation measures are not correlated with the earlier period's (1959–69) productivity growth: more highly regulated industries had a greater productivity slowdown than average because they did worse than average during the 1970's, not because they did better than average during the 1960's. Also, the coefficient on *PAOC* is generally smaller than 1 in magnitude, suggesting, in terms of equation (7), only a measurement effect on TFP calculation, without a real effect on productivity of noncompliance inputs.

Given some evidence for a regulation-productivity link, can we tell how important it is quantitatively? The R^2s from the regressions indicate what fraction of the variation in productivity growth across industries can be explained by the regulation measures. They tend to be small, indicating that regulation measures alone can only explain 4–5 percent of the variation across industries in TFP growth. Even with other factors included, only slightly over 10 percent of TFP variation is explained. This is due in large measure to the calculation of productivity as a residual: Input growth accounts for most of the variation in output growth rates, so the majority of the remaining variation in output growth is due to random disturbances.[9]

The estimated impact of regulation on productivity growth for the average industry is found by multiplying each regulation coefficient by the mean value of the regulation measure. In Table 3, model A4, the *OSHINS* coefficient (−.005) times the *OSHINS* mean (.54) yields a contribution to the TFP slowdown of .27 percentage points per year. This is 19 percent of the total slowdown of 1.46 percentage points. These measures are presented at the bottom of Table 3, where it can be seen that the regulation measures together account for a slowdown of .44 percentage points, somewhat over 30 percent of the average industry's productivity slowdown. The standard error of this total effect is only .15, so it is significant (due primarily to *OSHINS*).

There are a number of potential objections to these results, of which three will be treated here. First, the measures of regulation might themselves be affected by the industry's productivity growth, with this relationship being misinterpreted as an effect of regulation on productivity. Second, the linear regression model might be giving excessive weight to a few outlying industries with high regulation and poor productivity performance. Third, there could be some other explanation for the slowdown, not included in the current set of controls, whose omission biases the regulation coefficients.

[8] All of the standard errors are corrected for arbitrary heteroskedasticity, using the procedure suggested by Halbert White (1980). These corrections generally increased the standard errors of *PAOC* and decreased those of *OSHINS*.

[9] We can use equation (8) to represent TFP growth as the difference between output growth and aggregate input growth. Then *TFPCHG* is the difference between the change in output growth from the earlier to the later period (*OUTCHG*) and the change in aggregate input growth (*INCHG*). Regressing *OUTCHG* on *INCHG* along with the regulation measures and the other factors, we get

$$\begin{aligned} OUTCHG = &\underset{(.0051)}{.0082} + \underset{(.074)}{.864}{*}INCHG \\ &- \underset{(.0016)}{.0055}{*}OSHINS - \underset{(.52)}{.66}{*}PAOC, \end{aligned}$$

with an R^2 of .76 and standard errors in parentheses.

TABLE 4—REGRESSIONS RESULTS EXCLUDING OUTLIERS
(Subsample, $n = 439$)

Dependent Variable	Const.[a]	*OSHINS*	*PAOC*	*SHEN*	*SHCAP*	*GPLCHG*	*TFPCHGX*	$R^2(SSE)$
TFPCHG	.0100	−.0086	−0.70	−.063	−.060	.103	.133	.113
	(.0054)	(.0029)	(.71)	(.104)	(.019)	(.028)	(.050)	(.397)
TFP7378	.0102	−.0077	−0.83	−.084	−.028	.055	.101	.084
	(.0050)	(.0026)	(.55)	(.089)	(.018)	(.030)	(.049)	(.346)

Note: Standard errors are shown in parentheses.
[a]Subsample excludes 7 industries with *OSHINS* > 3.0 and 4 with *PAOC* > .025.

TABLE 5—REGRESSION RESULTS INCLUDING NONLINEARITY TEST
(Full sample, $n = 450$)

Dep. Var.	Const.[a]	*OSHINS*	*PAOC*	*OSH*PAOC*	$OSHINS^2$	$PAOC^2$	$R^2(SSE)$
TFPCHG	.0118	−.0136	−1.06	.053	.0023	18.88	.120
	(.0055)	(.0047)	(1.23)	(.260)	(.0012)	(38.9)	(.400)
TFP7378	.0113	−.0090	−1.33	−.017	.0012	29.05	.090
	(.0051)	(.0044)	(1.00)	(.217)	(.0011)	(32.2)	(.350)

Note: Standard errors are shown in parentheses.
[a]Both regressions also include *SHEN*, *SHCAP*, *GPLCHG*, and *TFPCHGX*.

The reverse causality argument, that industry productivity affects the regulation measures, could bias the results in either direction. Industries that were doing poorly (in productivity terms) might choose to reduce their spending on pollution abatement, leading *PAOC* to be underestimated.[10] On the other hand, if OSHA responded to diminished compliance expenditures by stepping up enforcement activities, the *OSHINS* coefficient would be overstated. Redoing all the equations using the 1973 values of *OSHINS* and *PAOC* (which should not be affected by post-1973 productivity growth) leaves the regression coefficients essentially unchanged. In a more formal test suggested by Hausman (1978), the predicted values of *OSHINS* and *PAOC* are insignificant when included in the regression, supporting the treatment of the regulation measures as exogenous.[11]

Possible failings of the linear regression model are examined next. Table 4 shows that when a few industries with exceptionally high regulation values are excluded from the regression, the coefficients on *PAOC* are almost unchanged, while the coefficients on *OSHINS* nearly double. This suggests that the marginal effect of OSHA regulation declines as regulation increases.[12] Table 5 tests directly for such nonlinear effects, showing that marginal impact declines (though not significantly) for both regulation measures. Table 6 presents a simple, nonparametric

[10]In terms of equation (7), θ_j is positively correlated with ε_j, biasing the estimate of β downward (in absolute magnitude).

[11]The variables used to explain *PAOC* and *OSHINS* are the cost share of labor, industry concentration and measures of establishment size within the industry, the 1973 regulation variables, the other four exogenous variables from Table 3, model A4, and for *OSHINS* the industry injury rate and an index of worker exposure to hazardous substances. These variables explain about 40 percent of the variation in *OSHINS* and 80 percent of the variation in *PAOC*. The results are available from the author.

[12]This result would seem to indicate that OSHA should do more inspections, since the inspections have declining marginal cost to the establishments inspected. This need not be true, since the marginal benefit of further inspections is also likely to decline as the number of inspections increases.

TABLE 6—MEAN TFP VALUES FOR INDUSTRY QUARTILES, RANKED BY EACH REGULATION MEASURE
(1 = lowest regulation, 4 = highest regulation)

	Ranked by *OSHINS*				Ranked by *PAOC*		
	Mean Values of:				Mean Values of:		
Q	*OSHINS*	*TFPCHG*	*TFP7378*	*Q*	*PAOC*	*TFPCHG*	*TFP7378*
1	.122	−.0072	.0009	1	.0004	−.0058	.0016
2	.256	−.0087	.0002	2	.0010	−.0133	−.0031
3	.461	−.0204	−.0104	3	.0019	−.0206	−.0106
4	1.325	−.0223	−.0125	4	.0082	−.0188	−.0095

examination of the regulation-productivity connection. Industries with high values for the regulation measures have slower productivity growth and a greater slowdown. In another nonparametric test, the Spearman rank correlations between *TFPCHG* and the regulation measures are negative and significant (−.19 for *OSHINS* and −.16 for *PAOC*). These tests do not suggest that the initial results are an artifact of the linear model.

The third objection, of omitted explanatory factors, is more difficult to address. A number of possible explanatory factors, including other input cost shares, the fraction of nonproduction workers in the work force, and changes in the use of various inputs, were tested and found not to affect regulation or productivity growth (these results, and all others referred to but not presented, are available from the author). When the four factors tested here are added to the regressions separately, the only one that affects the regulation coefficients noticeably is the energy cost share (*SHEN*), whose inclusion reduces the *PAOC* coefficient to −0.65. This could be due in part to multicollinearity, given the high correlation between *SHEN* and *PAOC*. The coefficient on *OSHINS* appears to be quite robust, although one cannot rule out the possibility of some other factor (as yet untested) being involved.

IV. Conclusions and Future Work

This study has found evidence that OSHA and EPA regulation reduced productivity growth in the average manufacturing industry by .44 percentage points per year, over 30 percent of the slowdown in the 1970's. This effect is larger than that found by most previous studies, which could be due to the focus on manufacturing, which faces relatively high OSHA and EPA regulation. The major surprise is that the effect of OSHA is quite strong, while that of EPA is comparatively weak. The prevailing opinion (that EPA has had more effect on productivity) may be due in part to the lack of good measures of compliance costs for OSHA regulation. There is also some evidence to indicate that pollution-abatement spending only affected the measurement of productivity growth, with no real effect on the productivity of inputs actually used in production. This result may be sensitive to the assumption of negligible regulation before 1970 and the strong ties between pollution-abatement spending and energy intensity. Further research is needed to resolve these issues.

I see two main areas for future research: collecting additional data and developing a more detailed model. Collecting more years of data will allow testing the effect of changing regulation over time, especially in recent years when the growth of regulation may have been reversed. Collecting variables measuring other factors such as research and development spending, labor quality, and market structure may help explain more of the slowdown. Finally, a model of the effects of regulation on productivity that explicitly incorporates a production or cost function will provide tests for whether the effect of regulation on productivity differs across inputs, and will allow a more complete explanation of why regulation affects productivity.

DATA APPENDIX

To calculate TFP growth, we need the real growth rate of output (measured here by the value of industry shipments) and the real growth rate and cost shares of each of five inputs: production workers, nonproduction workers, nonenergy materials, energy, and capital. The *Annual Survey of Manufactures* and the *Census of Manufactures*, published by the Census Bureau, provide data for each industry from 1958 to 1978 on the number of production worker hours, the number of nonproduction workers, and nominal measures of the value of shipments and expenditures on each input except capital. This enabled calculation of the cost shares for each input (capital's share is assumed to be 1 minus the sum of the other four inputs' shares).

Deflators for the value of shipments were provided by the Bureau of Industrial Economics in the Commerce Department. Various price indices from the Bureau of Labor Statistics were combined to create deflators for expenditures on energy and non-energy materials, with weights for energy based on 1976 expenditures on six types of energy, and weights for other materials based on the 1972 Input-Output tables.

The growth in the capital input for each industry is measured by the growth in the industry's real, depreciated capital stock. The capital stocks were initially calculated for the 1958–76 period by a joint project of the University of Pennsylvania, the Census Bureau, and SRI, Inc. They were extended by the author to 1978, based on data from the Bureau of Industrial Economics.

Further details on the procedures used can be found in my book (1986), and the data set itself is available upon request.

REFERENCES

Ashford, Nicholas A. and Heaton, George R., "Regulation and Technological Innovation in the Chemical Industry," *Law and Contemporary Problems*, Summer 1983, *46*, 109–57.

Christainsen, Gregory B. and Haveman, Robert H., "Public Regulations and the Slowdown in Productivity Growth," *American Economic Review Proceedings*, May 1981, *71*, 320–25.

Crandall, Robert W., "Pollution Controls and Productivity Growth in Basic Industries," in Thomas G. Cowing and Rodney E. Stevenson, *Productivity Measurement in Regulated Industries*, New York: Academic Press, 1981.

Darby, Michael R., "The U.S. Productivity Slowdown: A Case of Statistical Myopia," *American Economic Review*, June 1984, *74*, 301–22.

Denison, Edward P., *Accounting for Slower Economic Growth: The United States in the 1970s*, Washington: The Brookings Institution, 1979.

Gollop, Frank M. and Roberts, Mark J., "Environmental Regulations and Productivity Growth: The Case of Fossil-fueled Electric Power Generation," *Journal of Political Economy*, August 1983, *91*, 654–74.

Gray, Wayne B., *Productivity versus OSHA and EPA Regulations*, Ann Arbor: UMI Research Press, 1986.

Hoerger, Fred, Beamer, William H. and Hanson, James S., "The Cumulative Impact of Health, Environmental, and Safety Concerns on the Chemical Industry During the Seventies," *Law and Contemporary Problems*, Summer 1983, *46*, 109–57.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–71.

Jorgenson, Dale W., "The Role of Energy in Productivity Growth," *American Economic Review Proceedings*, May 1984, *74*, 26–30.

Norsworthy, J. R., Harper, Michael J. and Kunze, Kent, "The Slowdown in Productivity Growth: Analysis of Some Contributing Factors," *Brookings Papers on Economic Activity*, 2:1979, 387–421.

Portney, Paul R., "The Macroeconomic Impacts of Federal Environmental Regulation," in Henry M. Peskin, Paul R. Portney, and Allen V. Kneese, *Environmental Regulation and the U.S. Economy*, Baltimore: Johns Hopkins University Press, 1981.

Siegel, Robin, "Why Has Productivity Slowed Down?," *Data Resources Review*, March 1979, *1*, 1.59–1.65.

Viscusi, W. Kip, "Frameworks for Analyzing the Effects of Risk and Environmental Regulations on Productivity," *American Economic Review*, September 1983, *73*, 793–801.

Weisskopf, Thomas E., Bowles, Samuel and Gordon, David M., "Hearts and Minds: A Social Model of U.S. Productivity Growth," *Brookings Papers on Economic Activity*, 2:1983, 381–450.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48*, 817–38.

Deficit Announcements and Interest Rates

By PAUL WACHTEL AND JOHN YOUNG*

Despite the fact that the most theoretical analyses (with the notable exception of the Ricardian equivalence approach) indicate that increased deficits cause interest rates to rise, the empirical evidence is at best inconclusive.[1] In this note the relationship between interest rates and deficits is examined with the announcement effect methodology which has not previously been used in this context. We find evidence of a positive relationship between unanticipated announcements of the projected Federal government deficit and interest rates.

In an efficient market, information about any determinant of interest rates should be quickly incorporated into observed rates. Thus, when information about the size of the deficits is released, a relatively quick impact on interest rates can be anticipated. More specifically, if an increase in the deficit is, in fact, associated with higher interest rates, then an unanticipated announcement of a larger deficit should lead to a response in financial markets, which increases interest rates. This paper provides evidence on the announcement effects of information on the deficit. The advantage of the announcement effect approach is that it precludes the necessity of specifying a structural model for interest rates.[2]

Projections of current and future Federal government deficits are made on a regular basis by both the Office of Management and Budget (OMB) and the Congressional Budget Office (CBO), and receive wide attention in the financial press. These projections provide data that are related to the change in interest rate on government securities from the day before the announcement to the end of the announcement day.

The macroeconomic hypothesis underlying this investigation is simply that an increase in the current or future deficit leads to an increase in yields on government securities in anticipation of higher levels of debt financing. In a rational expectations framework, an announcement of higher future deficits will lead to a current increase in interest rates in anticipation of future financing. Thus, the examination of announcement effects enables us to substantiate a relationship between interest rates and deficits without encountering the econometric problems of reduced form modeling.

Section I begins with a description of the data. This is followed by a discussion of the methodology in Section II. Section III presents the empirical results. This is followed by our conclusions in Section IV.

I. Data

Deficit Announcements. Both the OMB and CBO make regular announcements of their projections of the deficit for the current fiscal year and for at least two fiscal years in the

*Professor of Economics, New York University Graduate School of Business Administration, New York, NY 10006, and Economist, Chase Manhattan Bank, New York, NY 10081, respectively. The authors are grateful to Laurence Ball, John Merrick, and George Sofianos for helpful comments.

[1] For example, recent papers by Paul Evans (1987) and Gregory Hoelscher (1986) reach opposite conclusions with somewhat similar quasi-reduced form equations for interest rates but with different data and time periods. Hoelscher concludes that larger deficits are associated with higher long-term rates, and Evans contends that there is no relationship with either short- or long-term rates. Further evidence in support of the Ricardian equivalence proposition is found in Charles Plosser (1982) who uses an efficient markets approach. He finds that innovations in the debt do not affect asset returns. Makin (1983) and Dewald (1983) find some weak effects of deficits on interest rates which are not large enough to account for the high level of interest rates in the early 1980's.

[2] The simultaneity problem makes it very difficult to identify the structural relationship between deficits (the supply of securities) and interest rates. The relationship is obscured by the fact that deficits are largely countercyclical (due to changes in tax revenues), while interest rates are procyclical (as private credit demands change).

future. The first such announcement is prepared by the OMB in late January or early February as part of the president's budget.[3] It provides a forecast for the fiscal year that began in October and for the next two fiscal years. These are sometimes fol lowed by revisions released in the spring. The CBO prepares its own projections based on its assumptions about the economy (output, inflation, and interest rates), which may or may not differ from the OMB assumptions. The CBO projections are released after the budget message and are always revised in late summer.

The definition of the deficit used here is the total budget deficit. It is the sum of the deficit on the unified budget and the off-budget deficit, which includes items such as funds raised by the Federal Financing Bank. It is the appropriate measure for three reasons. First, it reflects the total burden of federal deficits on the financial markets. Second, it is the most common measure referred to in press discussions of the deficit problem. Third, these total deficit projections are widely used by private forecasters as a benchmark in discussions of the deficit.

The data from OMB are available from 1979 to 1986; and the data from CBO start in 1982, when CBO started using the total budget deficit measure. The announcement dates and the projections for the current and two following fiscal years are shown in Table 1. The data set contains a relatively small number of announcements—22—spread over an eight-year period. An effort was made to augment the sample by scanning the *Wall Street Journal* index for additional announcements of projected deficits by both government and congressional sources. These included "leaks" from various officials, congressional testimony, and speeches. The augmented sample included over 50 announcements. However, these additional announcement observations did not change the results appreciably. Hence, results with the augmented data set are not presented.

TABLE 1—DEFICIT PROJECTIONS

	(Billions of Dollars) Fiscal Year		
Announcement Date	Current	Next ($T+1$)	Year Ahead ($T+2$)
Office of Management and Budget (OMB)			
1/22/79	49.4	41.0	12.7
1/28/80	56.5	33.9	10.3
3/31/80	51.5	2.2	N/A
1/15/81	78.4	45.8	23.4
3/10/81	78.5	61.7	N/A
2/8/82	118.3	107.2	106.3
1/31/83	224.8	202.8	204.7
2/1/84	199.9	195.2	185.9
2/4/85	222.2	180.0	164.9
2/5/86	202.8	143.6	93.6
Congressional Budget Office (CBO)			
9/10/81	84	85	N/A
2/8/82	129	176	206
9/2/82	130	173	170
2/4/83	210	212	231
8/22/83	223	208	195
2/7/84	203	208	230
8/7/84	183	191	209
2/6/85	214	215	233
2/27/85	215	220	240
8/15/85	210	212	229
2/18/86	208	181	165
8/7/86	224	184	150

Note: N/A means not available.

The efficient markets approach indicates that only new information should affect interest rates. Our preferred measure of the new information in an announcement of a deficit projection (or using terminology common in this context—the unanticipated part of the deficit projection) is simply the difference between the current projection and the previous projection for the same fiscal year. That is, let X_t^T be the deficit projection for fiscal year T announced at time t. The new information is defined as

$$X_t^T - X_{t'}^T,$$

where the previous announcement of X^T took place at t'. Thus, the unanticipated component of the deficit announcement is proxied by the revision in the projected deficit.

[3]These projections are found in the *Special Analyses, Budget of the U.S. Government* and in *Fiscal Year Budget Revisions*.

Since the amount of time between announcements (at t and t') can be as much as 6 months, the change in the projected deficit may reflect the effect of other information on the projected deficit, as well as the unanticipated part of the announcement. Therefore, an additional proxy for the unanticipated part of the deficit projection was constructed by regressing the change in the projection on economic data that became available between t and t'. The residuals from these regressions were then used as an alternative proxy for the unanticipated part of the change in the deficit projection. For the CBO announcements, this procedure resulted in announcement effects which were little different from those shown below with our preferred proxy for the unanticipated component. However, for the OMB announcements that occur less frequently, this procedure resulted in weaker results. The results in Section III used the preferred measure, the announced change in the deficit projection.

Interest Rates. The interest rate data are daily closing yields to maturity. They are the constant maturity yield series calculated by the Treasury and released by the Federal Reserve Board. Rates are available for 3-, 6-, and 12-month bills, and 2-, 3-, 5-, 7-, 10-, 20-, and 30-year notes and bonds. Each rate is a weighted average of the closing yields on various outstanding Treasury securities.

II. Methodology

The methodology employed here has been widely used elsewhere to examine various events. In particular the literature on money supply announcement effects uses this approach. The basic framework is given by

$$\Delta R = \alpha_0 + \alpha_1 (Z - Z^e),$$

where ΔR is the change in interest rates, Z is an information announcement, and Z^e is the expected announcement. Thus, $(Z - Z^e)$ is the unanticipated component or the new information, and ΔR is the change in interest rates from immediately before to immediately after the announcement. An announcement effect is present if α_1 is significantly different from zero.

It is possible that the financial markets also react to the anticipated part of the announcement, although this would indicate some inefficiency in the use of information. To test for this, equations of the following type are suggested:

$$\Delta R = \beta_0 + \beta_1 (Z - Z^e) + \beta_2 Z.$$

Market efficiency implies that $\beta_2 = 0$.

For the results in this paper, ΔR is the daily change in interest rates. The announcement information is a vector that is zero on non-announcement days and has the appropriate measure on announcement days.[4] Since the announcement effect for the weekly money supply has been documented in the literature, it is also included. Thus, the basic estimating equation is

$$(1) \qquad \Delta R_t = \gamma_0 + \gamma_1 (M_t - M_t^e) + \gamma_2 (X_t^T - X_{t'}^T),$$

where $(M - M^e)$ is the unanticipated component of the money supply announcement.[5]

Our interest in this paper is the γ_2 coefficients. A relationship between deficits and interest rates implies that there should be an announcement effect on interest rates. In particular, our hypothesis is that γ_2 is positive.

[4]The estimated regressions include all daily observations so that the effect of both deficit and money supply announcements which occasionally occur on the same day can be estimated. In this way, the estimated deficit announcement effects are not confounded by money supply announcement effects. However, this procedure leads to smaller coefficient standard errors than would be the case if the sample were restricted to deficit announcement days. Results with the smaller sample were not appreciably different from those shown below.

[5]The data are from the Money Market Services weekly surveys of the expected change in $M1$. Thomas Urich and Paul Wachtel, 1981 and 1984, develop the announcement effect methodology used here and examine the effects of money supply and price index announcements on interest rates.

III. Results

Tables 2 and 3 present the estimates of equation (1). The coefficients for the unanticipated money supply and deficit announcement effects are shown (the constant terms are omitted). The OMB and CBO deficit announcements are shown separately for two reasons. First, the data are available for different time periods. Second, the announcements often come at about the same time and embody similar information. Thus, when both are entered in the same equation, collinearity reduces the significance and size of the individual coefficients.[6]

The deficit announcements have a uniform positive effect on interest rates, as hypothesized. Although the OMB announcements precede the CBO announcements, the coefficients are larger for the CBO announcements.[7] In both cases, the announcement affects are about the same size all along the yield curve and are generally more significant for longer maturities.

A $1 billion change in the projected deficit leads to an average 0.30 basis point increase in interest rates for the CBO announcements and 0.18 basis points for the OMB announcements. These effects may be small because the measure of the unanticipated change in the deficit projection overestimates the actual surprise at the time of the announcement. It is likely to be an overestimate because it attributes the entire change from the previous projection usually made six months or a year earlier to an announcement day surprise. Much of this information would have already been widely known to financial markets participants and

[6] Considered jointly, the OMB and CBO deficit announcements are significant at the 10 percent level for half the interest rate maturities considered, somewhat less than the proportion of individual deficit announcement coefficients that are significant in Tables 2 and 3.

[7] The coefficients on the CBO announcements may be larger because the markets' participants have greater confidence in the economic assumptions made by CBO than those made by OMB. The CBO has a reputation for independence and integrity, while the OMB ends to reflect the political views and needs of the administration.

TABLE 2—CBO DEFICIT AND MONEY SUPPLY ANNOUNCEMENT EFFECTS[a]

Maturity of ΔR	Money Supply Effect	Deficit Effect	F^b
3-month bill	0.98	0.39	4.8
	(2.20)[c]	(2.19)	
6-month bill	1.03	0.36	2.4
	(1.67)	(1.46)	
12-month bill	0.74	0.33	4.6
	(2.00)	(2.25)	
2-year	0.96	0.27	4.6
	(2.50)	(1.73)	
3-year	1.11	0.32	5.8
	(2.75)	(2.00)	
5-year	0.80	0.30	3.8
	(2.02)	(1.88)	
7-year	0.70	0.31	4.1
	(1.91)	(2.14)	
10-year	0.70	0.29	4.1
	(2.00)	(2.03)	
20-year	0.63	0.35	5.0
	(1.84)	(2.56)	
30-year	0.55	0.36	5.5
	(1.74)	(2.83)	

Notes: [a] Sample period is 2/8/82 to 8/7/86; and sample size is 1,121.
[b] F statistic for regression.
[c] Coefficient t-statistics are in parentheses.

TABLE 3—OMB DEFICIT AND MONEY SUPPLY ANNOUNCEMENT EFFECTS[a]

Maturity of ΔR	Money Supply Effect	Deficit Effect	F^b
3-month bill	2.69	0.19	10.11
	(4.29)[c]	(1.35)	
6-month bill	2.25	0.19	9.27
	(4.04)	(1.52)	
12-month bill	2.12	0.15	10.95
	(4.48)	(1.37)	
2-year	2.15	0.20	11.29
	(4.38)	(1.85)	
3-year	2.05	0.20	11.36
	(4.39)	(1.88)	
5-year	1.85	0.17	10.29
	(4.21)	(1.71)	
7-year	1.58	0.14	9.06
	(3.91)	(1.68)	
10-year	1.58	0.14	9.92
	(4.14)	(1.65)	
20-year	1.55	0.18	11.27
	(4.21)	(2.21)	
30-year	1.22	0.18	8.77
	(3.49)	(2.32)	

Notes: [a] Sample period is 1/28/80 to 2/5/86; and sample size is 1,367.
[b] F statistic for regression.
[c] Coefficient t-statistics are in parentheses.

is likely to have affected interest rates prior to the deficit announcements. Nevertheless, there is a substantial impact of the announcements. The mean absolute change in the CBO deficit projections is $12 billion, which implies a 3.5 basis point change in interest rates on the announcement day. For the OMB projections, the mean absolute change is $36 billion, which implies a 6.5 basis point change in interest rates on the announcement day.

The statistical significance of these coefficients tends to increase as we move along the yield curve. In the CBO results, all but one of the deficit coefficients are significant at the 90 percent level and the deficit effects on 20- and 30-year bonds are significant at the 99 percent level. For the OMB announcements, the deficit effects on bill yields are not significantly different from zero at the 90 percent level and only the 20- and 30-year bond results are significant at the 95 percent level. The intermediate bond and note results are significant at the 90 percent level.

The important implication of the results in Tables 2 and 3 is that announcements concerning projected deficits are related to interest rates on government securities. In particular, longer-term yields increase when a deficit projection is increased. In fact, deficit projections were increased by as much as $50 to $100 billion in 1981–83. Such changes in projections could be associated with increases in yields of 9 to 30 basis points, depending on which estimates are used. Although not extremely large, these do provide firm evidence of a financial market response to information concerning deficits.

The results for the money supply announcements are consistent with those found in other studies. They are smaller for the CBO sample period that starts in 1981 and larger for the OMB period. The OMB sample period includes the period of the Fed's strongest commitment to unborrowed reserves targeting, which is the period when the money supply announcement effect was strongest. It declined when the Fed reverted to interest rate targets in late 1982 and virtually disappeared when contemporaneous reserve accounting was introduced in early 1984.

In contrast to the deficit announcements, the money supply announcement effects vary with maturity. They are strongest for short-term yields which are likely to be most seriously effected by changes in monetary policy. The money supply announcement effect on the long-term bond rate is only about half as large as the effect on bill yields.

The deficit measure used in Tables 2 and 3 is the change in the projected deficit for the current fiscal year. The OMB and CBO projections (see Table 1) include projections of the deficit for two subsequent fiscal years as well. Changes in projected future fiscal year deficits have a weaker effect on interest rates than the results shown. The difference is more pronounced for the CBO data than for the OMB data. Future deficits should have little effect on short-term securities that mature before the deficit is financed. The effects are smaller than for long-term securities, but even the current fiscal year deficit projections have little effect on bill yields.

The test for market efficiency with regard to information on the deficit involves adding the actual announcement to the equation. The coefficients for the actual deficit announcements are never significantly different from zero.

IV. Conclusion

In this note, we provide some striking evidence of the existence of an empirical link between interest rates and future government deficits. On the day that government agencies release information about projected Federal deficits, financial markets respond to these announcements. An increase in the projected deficit leads to an increase in interest rates. The announcement effect is felt on government securities of all maturities but is more significant on longer-term yields. The announcement effect methodology employed here does not provide any insight into the reasons for the observed relationships. As Plosser (1982) notes, increased deficits may increase interest rates because deficits crowd out real capital accumulation or because they will be monetized in the future or because they are due to government expenditure increases that affect

interest rates. Future research should attempt to distinguish among these three channels of the deficit effect on interest rates.

REFERENCES

Dewald, William G., "Federal Deficits and Real Interest Rates: Theory and Evidence," *Federal Reserve Bank of Atlanta Economic Review*, January 1983, *68*, 20–29.

Evans, Paul, "Interest Rates and Expected Future Deficits in the United States," *Journal of Political Economy*, February 1987, *95*, 34–58.

Hoelscher, Gregory, "New Evidence on Deficits and Interest Rates," *Journal of Money, Credit, and Banking*, February 1986, *18*, 1–17.

Makin, John H., "Real Interest Rates, Money Surprises, Anticipated Inflation and Fiscal Deficits," *Review of Economics and Statistics*, August 1983, *65*, 374–84.

Plosser, Charles I., "Government Financing Decisions and Asset Returns," *Journal of Monetary Economics*, May 1982, *9*, 325–52.

Urich, Thomas and Wachtel, Paul, "Market Response to the Weekly Money Supply Announcement in the 1970s, *Journal of Finance*, December 1981, *36*, 1063–72.

______ **and** ______, "The Effects of Inflation and Money Supply Announcements on Interest Rates," *Journal of Finance*, September 1984, *39*, 1177–88.

Executive Office of the President, Office of Management and Budget, *Special Analyses, Budget of the U.S. Government*, Washington, D.C.: USGPO, various years.

The Price of Final Product After Vertical Integration

By SEUNG HOON LEE*

One of the main strands in the studies of vertical integration investigates the behavior of final product price when the forward integration by the upstream monopolist leads to a monopoly in the market for downstream product. So far the price of final product is known to remain the same as before if the production technology of downstream industry is using inputs only in fixed proportions. As in the case of variable proportions technology, the most up-to-date results had been reported by Parthasaradhi Mallela and Babu Nahata (1980), and Fred Westfield (1980). Mallela and Nahata studied the case when the downstream technology is representable by a constant elasticity of substitution (CES) production function and the demand for final product exhibits a constant price elasticity. They found that the price of the final product must rise after integration when the substitution elasticity has a value not less than unity. On the other hand, Westfield presented an alternative analytic framework by introducing the concept of *benchmark* price of the intermediate good, which will sustain the (integrated merger's) monopoly equilibrium configuration of final product market as a competitive configuration before integration as well. But there is an inherent problem in Westfield's approach: All of his criteria utilize the values of parameters evaluated at the configuration of not current but benchmark price, and therefore those values are not to be observed directly or computed easily from the observation of the real world. In the benchmark configuration, the downstream market exhibits the same equilibrium price and quantity as those in monopoly which would be induced by the vertical merger. If one knows precisely what the benchmark configuration, especially the price of final product, will be, then no further analysis will be necessary!

There is another aspect of this problem that has never been addressed to so far. All the pre-existing results assume a constant upstream marginal cost, excluding many standard cases where the upstream monopoly is a natural one. Thus it will be worth investigating what happens when the upstream marginal cost is variable. In this note, I propose an alternative set of criteria which will generalize and improve the pre-existing ones by comparing the marginal revenue of downstream industry directly with the marginal cost of vertical merger at the current level of production before integration. Although this marginal cost might not be observable, this comparison yields the conditions consisting of parameters whose values are observable from the current configuration. Assume that the demand for final product exhibits a nonincreasing marginal revenue schedule. Then whenever the vertical integration yields a marginal cost for producing the final product which is higher (lower) than the marginal revenue at the current level of production, the integration with monopolization will surely lead to a decrease (an increase) of quantities produced and the corresponding increase (decrease) in the price of final product. All the analyses below will be carried out according to this criterion.

I will investigate the conventional case where the final product is produced from one primary factor and one intermediate factor. The primary factor is sold at a constant competitive price that is set at unity and the intermediate factor is sold by the upstream monopolist at the price s. Assume that the technology of downstream industry is representable by a linearly homogeneous neoclassical production function. Then the down-

*Associate Professor of Economics, Department of Economics, Seoul National University, Seoul 151, Korea.
I appreciate the comments of A. Mas-Collel, D. Abreu, W. Thomson, and an anonymous referee. I also gratefully acknowledge the fellowship support from the Harvard-Yenching Institute.

stream marginal cost will depend only on the factor price s and it will be denoted by $\lambda(s)$. Notice that this is also equal to the competitive price of the final product. Denote r for the upstream marginal cost, ε for the price elasticity of demand for the final product, σ for the elasticity of substitution in the downstream technology, and β for the relative cost share of intermediate input, respectively. All these notations will represent the values evaluated at the current observation before integration, unless indicated otherwise.

After integration, the vertical merger's marginal cost for producing an additional unit of the final product will become $\lambda(r')$. The r' indicates the new level of upstream marginal cost which will be determined by the amount of intermediate input used by the vertical merger. Since $\lambda(s)(1-1/\varepsilon)$ is the marginal revenue that will be faced by the vertical merger at the current configuration of downstream market, the vertical integration will raise the price of the final product if and only if $\lambda(r') > \lambda(s)(1-1/\varepsilon)$. One immediate case is $\varepsilon \leqq 1$ as long as $\lambda(r') > 0$. This occurs only if $\sigma > 1$ since $\beta\varepsilon + (1-\beta)\sigma > 1$ (see below). When $\varepsilon > 1$ prevails, I will assume that the upstream marginal cost is nondecreasing. Then $\lambda(r) > \lambda(s)(1-1/\varepsilon)$ will be sufficient for the price of final product to rise. It can be shown that the upstream monopolist sets his monopoly price s so that

$$s\left(1-\frac{1}{\beta\varepsilon+(1-\beta)\sigma}\right)=r$$

holds with $\beta\varepsilon+(1-\beta)\sigma > 1$. Now one has

$$\lambda(r)=\lambda\left(s\left(1-\frac{1}{\beta\varepsilon+(1-\beta)\sigma}\right)\right)$$
$$>\lambda(s)\left(1-\frac{1}{\beta\varepsilon+(1-\beta)\sigma}\right)$$

since the right-hand side is the cost for producing unit output when the prices of all inputs are lowered proportionately from $(1, s)$ (remember that λ is linearly homogeneous in all input prices), while the left-hand side is the same cost when only the price of intermediate input is lowered at the same proportion. From this inequality one can derive a sufficient condition for the price to rise as

$$\lambda(s)(1-1/[\beta\varepsilon+(1-\beta)\sigma])$$
$$\geqq \lambda(s)(1-1/\varepsilon),$$

or equivalently, $\varepsilon \leqq \sigma$.

Now consider the case $0 < \sigma \leqq 1$. This case constitutes a part of the case $\varepsilon > \sigma$, which has been left out so far. Observe that, for the price t of intermediate input, the differential equation

$$\frac{t}{\lambda(t)}\frac{\partial\lambda(t)}{\partial t}=\frac{tb(t)}{\lambda(t)}=\beta(t)$$

holds, where $b(t)$ and $\beta(t)$ denote the input coefficient and the relative cost share of intermediate input, respectively, when its price is t (use Shepherd's Lemma). Suppose the elasticity of substitution for the downstream technology is, although not necessarily constant, not greater than 1 at *every* input combination. Then $\beta(t) \leqq \beta$ holds for all t with $r \leqq t \leqq s$ and therefore one obtains

$$\lambda(r) \geqq \lambda(s)\left(\frac{r}{s}\right)^{\beta}$$
$$=\lambda(s)\left(1-\frac{1}{\beta\varepsilon+(1-\beta)\sigma}\right)^{\beta}$$

by integrating the above differential equation. Define $k(\tau) \equiv (1-1/[\beta\varepsilon+(1-\beta)\tau])^{\beta}$. It is easy to see that $k(\tau)$ is strictly increasing in τ for each set of fixed values β and ε when $\tau > -\beta\varepsilon/(1-\beta)$ holds, and in turn that there exists a real number σ^* with $\varepsilon > \sigma^*$ so that $(1-1/[\beta\varepsilon+(1-\beta)\sigma^*])^{\beta} = 1-1/\varepsilon$ holds. This number σ^* is solved as

$$\sigma^*=\frac{1}{\left[(1-1/\varepsilon)^{1/\beta}-1\right](\beta-1)}+\frac{\beta}{\beta-1}\varepsilon,$$

and clearly can always be computed from the knowledge of the values of β and ε. Now let

$$h(u) \equiv (1-1/u)^{(u-1)/(\varepsilon-1)}.$$

Now we have $h(\beta\varepsilon+(1-\beta)) = k(1)$. Since one has (use $\ln(1-x) < -x$ for $0 < x < 1$)

$$\frac{\partial \ln h(u)}{\partial u} = \frac{1}{\varepsilon-1}\left[\frac{1}{u} + \ln\left(1-\frac{1}{u}\right)\right] < \frac{1}{\varepsilon-1}\left(\frac{1}{u}-\frac{1}{u}\right) = 0,$$

and since $h(\varepsilon) = 1-1/\varepsilon$ holds with $\varepsilon > \beta\varepsilon + (1-\beta)$, it is clear that $k(1) > 1-1/\varepsilon$. This in turn implies $\sigma^* < 1$. Therefore, if $\sigma^* < \sigma \leqq 1$ holds for the currently observed values of β and ε, and if the value of elasticity of substitution, although not necessarily constant, is everywhere not greater than 1, then the price of final product will necessarily rise after the integration. Notice that this conclusion accounts for the case of Cobb-Douglas production function, too.

Now let us generalize the Mallela-Nahata result to the case of variable price elasticity of demand for the final product. It is sufficient to consider only the case of $1 < \sigma < \varepsilon$. Consider the CES production function[1]

$$q = \gamma[\delta x^{\rho} + (1-\delta)y^{\rho}]^{1/\rho}, \qquad \gamma > 0,\ 0 < \delta < 1,\ \rho < 1$$

and the associated marginal cost at the price t of intermediate input

$$\lambda(t) = \frac{\delta^{\sigma/(1-\sigma)}}{\gamma}\left[\left(\frac{1-\delta}{\delta}\right)^{\sigma} t^{1-\sigma} + 1\right]^{1/(1-\sigma)}.$$

Let $e \equiv \beta\varepsilon + (1-\beta)\sigma$. It can be shown that

$$\beta = 1\bigg/\left[\left(\frac{1-\delta}{\delta}\right)^{-\sigma} s^{\sigma-1} + 1\right],$$

and thus

$$\frac{e-\sigma}{\varepsilon-e} = \frac{\beta}{1-\beta} = \left(\frac{1-\delta}{\delta}\right)^{\sigma} s^{1-\sigma}.$$

Therefore, one has

$$\frac{\lambda(s)(1-1/\varepsilon)}{\lambda(r)} = \left[\frac{\dfrac{\varepsilon-\sigma}{\varepsilon-e}(1-1/\varepsilon)^{1-\sigma}}{\dfrac{e-\sigma}{\varepsilon-e}(1-1/e)^{1-\sigma}+1}\right]^{1/(1-\sigma)}.$$

When $\varepsilon > e > \sigma > 1$ prevails, it holds without assuming the constant price elasticity of demand for the final product that

$$\lambda(s)\left(1-\frac{1}{\varepsilon}\right) \gtreqless \lambda(r)$$

$$\text{iff } (\varepsilon-\sigma)\left[\left(1-\frac{1}{\varepsilon}\right)^{1-\sigma} - 1\right] \lesseqgtr (e-\sigma)\left[\left(1-\frac{1}{e}\right)^{1-\sigma} - 1\right].$$

Define

$$f(v) \equiv (v-\sigma)\left[(1-1/v)^{1-\sigma} - 1\right].$$

Now it is sufficient to show that $f(v)$ is strictly increasing in v whenever $\sigma > 1$ holds, since we already have $\varepsilon > e$. Since one has (use the inequality $(1-x)^a > 1-ax$ for $0 < x < 1$ and $a < 0$ or $a > 1$)

$$(1-1/v)^{1-\sigma} > 1-(1-\sigma)/v$$

and

$$(1-1/v)^{\sigma} > 1-\sigma/v$$

for every v with $v > 1$ when $\sigma > 1$ holds, we

[1] Here the output, and primary factor, and the intermediate input are denoted by q, x, and y, respectively.

obtain

$$\frac{\partial f(v)}{\partial v}$$

$$> -\frac{1-\sigma}{v}\left[1-(1-1/v)^{\sigma}(1-1/v)^{-\sigma}\right]=0,$$

and this completes the proof.

Now the results that I obtained so far in this analysis can be summarized as follows: Assume a downward-sloping marginal revenue schedule for the final product. Also assume that the downstream technology exhibits the constant returns to scale and the upstream technology exhibits a nondecreasing marginal cost. Then it is shown that (*i*) that if the currently observed value of price elasticity of demand for the final product is not larger than that of the elasticity of substitution for the technology producing final product, then the vertical integration will always lead to an increase in the price of final product via monopolization; (*ii*) that there exits a critical number σ^*, whose value is always computable from the current observation and is always smaller than 1, so that if the elasticity of substitution has a value which is larger than this critical number at the current observation before integration and which is everywhere not larger than 1, then the price of final product will rise after integration; and (*iii*) that if the downstream technology is representable by any CES production function with the value of substitution elasticity which is not smaller than 1, then the monopoly price of final product after integration will be higher than the competitive price before integration even when the price elasticity of demand for final product is variable.

If one adds to these the result for the case of fixed proportions technology (which has been known since early days), then a most comprehensive set of sufficient conditions will be obtained for the monopoly price of final product after integration to be higher than the competitive price before integration. And it should be clear that the above conditions, except that for the fixed-proportions technologies, should not be applied to the cases where the upstream monopoly is a natural one with the strictly decreasing marginal cost. It is because in these cases the increase in use of the intermediate input after integration will lower the marginal cost for producing it, and in turn will make the merger's marginal cost for producing the current level of final product lower than $\lambda(r)$. So even when the above conditions are satisfied, it may be the case that the new marginal cost is lower than the marginal revenue at the current level of production, and therefore, the price must fall.

REFERENCES

Hicks, J. R., *The Theory of Wages*, New York, 1968.

Mallela, P. and Nahata, B., "Theory of Vertical Control with Variable Proportions," *Journal of Political Economy*, October 1980, *88*, 1009–25.

Westfield, F. M., "Vertical Integration: Does Product Price Rise or Fall?," *American Economic Review*, June 1981, *71*, 334–46.

Industrialization, Deindustrialization, and North-South Trade

By ANDRÉ BURGSTALLER*

A straightforward extension of David Ricardo's corn model[1] to a North-South system[2] is capable of generating a richly allusive account of trade-induced industrial growth and decline. Consider the following five-act scenario.[3]

I

England, a flexible-wage, full-employment northern economy with a fixed labor force and land endowment, is producing the wage-good corn (using labor and land) and two luxury manufactures consumed out of rent, textiles, and motorcars (using labor only). India, a southern fixed-wage Malthus-Marx-Lewis labor-surplus economy, also is producing corn on a fixed amount of capitalistically usable land, but as yet no manufactures. Profit rates are positive in both countries and profits fully saved and locally invested. From the postulated labor market asymmetry, it is clear that the process of accumulation of corn-wage funds by capitalists must have very different consequences in the two countries. Whereas in England it serves to raise the wage rate of a constant labor force, in India it draws on the available fixed-wage surplus labor to increase market employment. Under the impetus of the consequent growth in Indian corn rent and a correspondingly increasing Indian import demand for manufactures, the world market corn price of textiles and motorcars is rising, supporting a secular shift of England's resources out of agriculture into manufacturing. In short, England is engaged in a process of export-led industrialization.

II

The corn price of textiles has risen sufficiently to make their production in India profitable despite a relatively backward Indian textile technology. At given prices, any increment to the Indian capital stock will henceforth shun agriculture and move into textiles. The consequent rise in Indian textile output brings about repeated situations of excess supply in the world market for textiles. This continuously forces down the corn price of textiles, as well as that of motorcars, since the latters' textile price is locked in by the two constant English manufacturing labor productivities. Clearly, England's textile industry will be plunged into crisis by this development, losing some of its employment to agriculture, and some to a motorcar industry that must be booming since equilibrium world demand for motorcars is increasing (Indian and English corn rents are rising and the corn price of motorcars is falling).

III

English textiles have been wiped out by their Indian competition—the end result of a process that has endogenously shifted the basis for trade in textiles from differences in technology to differences in relative factor abundance. With England's industry now fully specialized in motorcars and India's in textiles, the textile price of motorcars is free to vary on world markets. Its course under continued capital accumulation is easy to discern: As previously, increases in India's capital stock must depress the corn price of textiles, thereby ensuring growth of Indian agriculture, corn rents, and import demand

*Department of Economics, Barnard College, Columbia University, New York, NY 10027.

[1] *Essay on Profits* (Ricardo, 1951, IV). We adopt Pasinetti's 1960 interpretive structure.

[2] See Findlay, 1980.

[3] Mathematical proofs are straightforward and omitted.

for motorcars; accordingly, the corn- and (a fortiori) textile price of motorcars must be rising. England's manufacturing sector, having formerly suffered an overall contraction vis-à-vis agriculture because of the collapse of textiles, therefore is enjoying a renaissance fueled by the continued, now purely export-led expansion of its motorcar industry. The fate of India's manufacturing sector, in contrast, is ambiguous: Though world demand for textiles is increasing on account of the fall in their corn price and the rise in Indian corn rents, it will be adversely affected by the decline of corn-rent income in England. If propensities to import manufactures are sufficiently high, the latter influence may dominate and enforce a prolonged contraction of Indian manufacturing[4] that will only come to a halt if and when India emerges as a competitive producer of motorcars.

IV

The continued rise in the corn price of the latter in fact makes it likely that India will so emerge. Once this happens, the textile price of motorcars again becomes invariant, fixed now by the constant labor productivities prevailing in Indian textile and motorcar production. Since the corn price of textiles continues to fall, so must the corn price of motorcars and for the same reason—accumulation-induced output expansions of India's automotive industry that repeatedly lead to a world excess supply of motorcars. The implications for English manufacturing are simple and severe: Its motorcars are progressively displaced in world markets by the Indian competition.

[4]See my paper, 1987, which examines stage III sectoral and overall employment effects under alternative interregional capital mobility regimes.

V

Ultimately, world manufacturing prices have fallen so low that the remnants of the English motorcar industry are wiped out by imports. England's process of deindustrialization has run its course—the formerly dominant industrial power has reverted to a purely agricultural economy that, but for remaining flickers of the redistributive struggle between workers and capitalists, is quiescent. Meanwhile, India's industry and agriculture are flourishing and expanding, albeit at a decelerating rate.[5]

[5]Note the following qualifications: (*i*) the fixed-wage assumption for India, maintained throughout the analysis, may in fact cease to hold at some point during the growth process; (*ii*) an early onset of the stationary state may bring the system to a halt before stage V is reached.

REFERENCES

Burgstaller, A., "Europe's Industrialization and Colonial Underdevelopment in the Light of Ricardo's Corn Model," *Journal of International Economics*, 1987, *22*, 157–69.

Findlay, R., "The Terms of Trade and Equilibrium Growth in the World Economy," *American Economic Review*, June 1980, *70*, 291–99.

Pasinetti, L., "A Mathematical Formulation of the Ricardian System," *Review of Economic Studies*, February 1960, *27*, 78–98.

Ricardo, D., *The Works and Correspondence of David Ricardo*, Vols. I–X, P. Sraffa, ed., Cambridge: Cambridge University Press, 1951.

The Public Finance of a Protective Tariff: The Case of an Oil Import Fee

By DAVID S. BIZER AND CHARLES STUART*

Recent debate has focused on the desirability of imposing an oil import fee or some broader tax on oil consumption in order to finance tax reform or for other purposes. To study the issue, we note that optimal taxation requires that the government raise revenue using the tax instrument with the lowest efficiency cost per dollar of additional revenue (Peter Diamond and Daniel McFadden, 1974). We use a highly stylized but conventional general-equilibrium model to evaluate the magnitude of this "marginal efficiency cost" for taxes on oil imports, oil consumption, and, as a reference for comparison, labor income. In addition to shedding light on the debate, the analysis provides several more general insights about the public finance of a protective tariff.

We find that the tax on labor income has the lowest marginal efficiency cost and the oil import fee has the highest. Of particular interest is that the import fee, as a protective tariff, can be viewed as a tax on oil consumption combined with off-budget lump-sum transfers to owners of domestic oil. For a given tax-induced rise in the price of oil to domestic consumers, the import fee thus raises less net revenue and hence has a greater marginal efficiency cost than the oil consumption tax.

Although a protective tariff is a tax that raises revenue and distorts resource allocation, application of the theory of optimal taxation is complicated by the fact that a tariff may export efficiency cost to foreign households. Determination of the optimal use of a tariff therefore requires modeling equilibria in a game of trade policy played by different countries. Full analysis of such equilibria is outside the scope of the present paper, but we examine two interesting cases. In the first, the rest of the world is a price-taking Nash player whose behavior is described by an upward-sloping supply curve. Here, an increase in U.S. tariffs improves U.S. terms of trade. In the second, the rest of the world retaliates. Although retaliation could in principle take many forms, we consider a strategy of levying tariffs against U.S. exports to hold terms of trade constant. Such an exercise is of interest because recent work indicates that tit for tat may arise as an equilibrium strategy in noncooperative extensive-form games (Robert Anderson, 1985), as well as in experimental games (Robert Axelrod, 1984). A protective tariff may have a particularly high marginal efficiency cost when the rest of the world retaliates.

I. Model

Consider an open economy in which a representative domestic household consumes leisure, oil, and an aggregate numeraire good.[1] The household derives income from sales of labor and domestic oil, ownership of capital, and lump-sum transfers from the government. A representative firm converts labor and capital into numeraire under competitive conditions. A government levies taxes on income, imported oil, and oil consumption (such as taxes on gasoline), and exhausts the resulting receipts on lump-sum transfers to the household and on government consumption.

In more detail, the price of oil obtained by domestic suppliers is $p_f + \phi$, where p_f is the

*Department of Economics, Stanford University, Stanford, CA 94305, and Department of Economics, University of California, Santa Barbara, CA 93106. We wish to thank Marty Zimmerman, two referees, and the Bank of Sweden Tercentenary Foundation.

[1] We include oil in utility for simplicity and because much oil is consumed relatively directly—for example, gasoline and heating oil.

foreign price and ϕ is the oil import fee. We allow for a supply response whereby p_f rises as imports of foreign oil (O_f) increase. The price of oil to domestic consumers, which includes oil consumption taxes at rate γ, is

$$(1) \qquad p = p_f + \phi + \gamma.$$

Total oil consumed (O) is domestically supplied oil (O_d) plus imported oil: $O = O_d + O_f$. For simplicity, we assume that O_d is constant.[2] Let C denote consumption of numeraire, τ denote the tax rate on labor income, w denote the wage, K denote capital income, and R denote lump-sum transfers from the government.[3] The household's budget is then

$$(2) \quad C + pO = (1-\tau)wL + (p_f + \phi)O_d + K + R.$$

Similarly, the government's budget is

$$(3) \qquad R + G = \tau wL + \phi O_f + \gamma O,$$

where G is government consumption. Together, (2) and (3) yield the national income identity net of consumption and production of domestic oil: $C + G + p_f O_f = wL + K$.

We assume that G is separable in the household's utility. The household chooses C, O, and L to maximize utility, $U(C, O, L, G)$, subject to (2), taking p, τ, w, p_f, ϕ, O_d, K, and R as given.

We take the aggregate production function to be concave and assume that capital is constant.

II. Welfare Costs

We measure welfare changes as the Hicksian compensating surplus, which is the root Z of $U(C', O', L', G') = U(C + Z, O, L, G)$, where primed (unprimed) variables denote equilibrium values before (after) a small rise in a tax instrument. In most cases, lump-sum redistributions are constant so the change in tax revenue equals the change in government consumption, $G - G'$. Marginal efficiency cost is then calculated as $Z/(G\text{-}G') - 1$.

The marginal efficiency cost for a change in the oil import fee typically exceeds the marginal efficiency cost for a change in the oil consumption tax here because the base of consumption tax is O but the base of the import fee is only O_f. The difference, O_d, represents an off-budget transfer to owners of domestic oil. That is, an oil import fee, as a protective tariff, taxes all domestic consumption but diverts some of the revenue gain to owners of the protected factor.[4]

III. Numerical Assessment

To assess magnitudes, we take utility to be generalized CES, $C^{-e} + b_0(O - d_0)^{-e} + b_1(d_1 - L)^{-e} + V(G)$, where b_0, d_0, b_1, d_1, and e are parameters and V is a function. Production is Cobb-Douglas, AL^a, where A and a are parameters with capital subsumed in A. The gross wage is $w = aAL^{a-1}$. The supply price of foreign oil is a constant-elasticity function, $p_f = \varepsilon_0 O_f^{\varepsilon}$, where ε_0 and ε are parameters. When the rest of the world plays a price-taking strategy, we chose ε to be the elasticity of oil prices to the United States under competitive conditions. When the rest of the world plays tit for tat to hold terms of trade constant, we simply take $\varepsilon = 0$.

To use this structure, we begin with a benchmark equilibrium that reflects early 1986 values for all real variables and tax rates, that is, $AL^a = 4000$ billion dollars/year, $a = .65$, $G = 400$ billion dollars/year, $O = 5.95$ billion bbl/year, $O_d = 4.09$ billion bbl/year, $O_f = 1.86$ billion bbl/year, $p_f = 15$

[2] This avoids the complication of modeling the supply response of domestic oil. The actual supply elasticity for domestic oil is small, perhaps around .2 (S. Fred Singer, 1983). Small additional distortions should therefore arise because the oil import fee causes relatively costly domestic production to rise.

[3] We treat capital income as untaxed without loss of generality because the sum $K + R$ would not change if capital were taxed explicitly.

[4] To see the intuition, perturb the initial equilibrium with a rise in ϕ of $\Delta\phi$. From (3), government consumption becomes $G_\phi \equiv \tau wL + (\phi + \Delta\phi)O_f + \gamma O - R$. Now perturb the initial equilibrium with an equal rise in γ instead. If private behavior were the same, government consumption would be $G_\gamma \equiv \tau wL + \phi O_f + (\gamma + \Delta\gamma)O - R$. Thus $G_\gamma - G_\phi = O_d \Delta\phi > 0$, so the rise in γ would yield more revenue and hence G than the rise in ϕ.

TABLE 1—EFFICIENCY COSTS PER DOLLAR OF ADDITIONAL TAX REVENUE

Case	Rise in				
	Oil Import Fee		Oil Consumption Tax		Labor Income Tax
	0 to $5/bbl	$5 to $6/bbl	$4 to $9/bbl	$9 to $10/bbl	.40 to .405
Basic					
1) Price Taking	1.36	170	.09	.19	.07
2) Tit for Tat	2.64	–	.17	.20	.07
Oil Elasticity = –.8					
3) Price Taking	4.78	–	.16	.25	.07
4) Tit for Tat	–	–	.29	.39	.07
Marginal Spending on *R*					
5) Price Taking	.48	26	.18	.24	.16
6) Tit for Tat	1.83	–	.25	.32	.16

Notes: No marginal efficiency cost is reported when a tax increase reduces revenue. Unless otherwise noted, the price elasticity of oil demand is $-.6$ and the uncompensated and compensated wage elasticities of labor supply are .10 and .25, respectively.

dollars/bbl, $\tau = .4$, $\gamma = 4$ dollars/bbl, and $\phi = 0$.[5] We assume that when the rest of the world is a price taker, the elasticity of p_f with respect to O_f is $\varepsilon = 0.103$, and we find ε_0 such that $p_f = 15$ when $O_f = 1.86$.[6] Under tit for tat, we take $\varepsilon = 0$ and $\varepsilon_0 = 15$. Utility parameters are derived by assuming that the price elasticity of domestic demand for oil is $-.6$ and the uncompensated and compensated wage elasticities of labor supply are .1 and .25, respectively, in the benchmark.[7]

IV. Results

A $5/bbl oil import fee is at the low end of recent proposals. We calculate marginal efficiency cost for a fee of this size and for an incremental increase in the fee from $5/bbl to $6/bbl. Similarly, we consider increasing the oil consumption tax from $4/bbl to $9/bbl and also report the marginal efficiency cost for an incremental rise in γ from $9/bbl to $10/bbl. The increase in the labor tax is incremental, from .40 to .405, in all cases.

The basic results are in rows one and two of Table 1. The oil import fee has a substantially greater marginal efficiency cost than either an oil consumption tax or a tax on labor income, especially when one considers increases from a level of $\phi = \$5$/bbl. Small increases in the import fee generate little additional tax revenue and hence have high marginal efficiency costs here because resulting decreases in the demand for oil are borne entirely by oil imports. For instance, the marginal efficiency cost of 170 in row one indicates that total tax receipts are a maximum when the import fee is roughly $6/bbl in this case. The import fee and to a lesser extent the tax on oil consumption have particularly high marginal efficiency costs when the rest of the world plays tit for tat.

Rows three and four report sensitivity analysis based on a price elasticity of demand for oil of $-.8$ instead of $-.6$. This makes taxes on oil imports and consumption less attractive relative to taxes on labor.

Finally, rows five and six contain marginal efficiency costs calculated under the assumption that all marginal public revenue is redis-

[5] Estimates of a, G, and τ are from Charles Stuart (1984). Taxes on oil consumption are primarily from levies on gasoline and diesel fuel, which average about 20¢/gallon. Because about half of the 42 gallons in a barrel of oil is consumed as motor fuel, γ is about 4 dollars/bbl. Data on O, O_d, and O_f are from the U.S. Department of Energy (1986).

[6] To find ε, we note that $O_f(p_f) = S_f(p_f) - D_f(p_f)$, where S_f and D_f are the total supply and demand functions in the rest of the world. We differentiate and invert this expression, put it into elasticity form, and assume the supply and demand elasticities are .2 and $-.6$.

[7] See Robert Pindyck, 1979; Energy Modeling Forum, 1982; Singer, 1983; and Ingemar Hansson and Stuart, 1985.

tributed as a lump sum to the domestic household. Stuart (1984) provides evidence that increases in Federal revenue in recent years have largely been redistributed. Marginal efficiency cost is $Z/(R-R')$ in this case.[8]

V. Conclusions

Analysis of the welfare effects of a protective tariff requires assumptions about the extent of foreign retaliation. We consider two cases: no retaliation and retaliation to hold terms of trade constant. Even without foreign retaliation, an oil import fee is substantially less efficient than a tax on labor income as a means of raising marginal public revenue. Because an import fee is equivalent to a tax on oil consumption combined with off-budget transfers to owners of domestic oil, the import fee is also less efficient than a tax on oil consumption for financing on-budget public spending.

In the cases we consider, the oil consumption tax is less efficient than a tax on labor income for raising marginal public revenue. However, the efficiency difference between the two is small.

Finally, we would emphasize that our calculations of efficiency cost per additional dollar of tax revenue neglect strategic, environmental, and other public objectives sometimes cited as justifications for energy taxes. If there is a strong efficiency case for such taxes, it presumably rests on factors not considered here.

[8] Stuart (1984) explains the formula. If providing transfers to owners of domestic oil is a policy objective, one may wish to include the transfer as part of the revenue increase. In this case, marginal efficiency costs for ϕ in rows five and six would be identical to those reported in columns three and four for an increase in γ.

REFERENCES

Anderson, Robert, "Quick-Response Equilibrium," unpublished manuscript, 1985.

Axelrod, Robert, *The Evolution of Cooperation*, New York: Basic Books, 1984.

Diamond, Peter and McFadden, Daniel, Some Uses of the Expenditure Function in Public Finance," *Journal of Public Economics*, February 1974, *3*, 3–21.

Hansson, Ingemar and Stuart, Charles, "Tax Revenue and the Marginal Cost of Public Funds in Sweden," *Journal of Public Economics*, August 1985, *27*, 331–53.

Pindyck, Robert S., *The Structure of World Energy Demand*, Cambridge: MIT Press, 1979.

Singer, S. Fred, "The Price of World Oil," *Annual Review of Energy*, 1983, *8*, 451–508.

Stuart, Charles, "Welfare Costs per Dollar of Additional Tax Revenue in the United States," *American Economic Review*, June 1984, *74*, 352–62.

U.S. Department of Energy, *Monthly Energy Review*, December 1985, Washington, D.C.: USGPO, 1986.

Energy Modeling Forum, *World Oil*, EMF Report 6, Stanford, CA: 1982.

Secession and the Limits of Taxation: Toward a Theory of Internal Exit

By James M. Buchanan and Roger L. Faith*

Since the publication of Charles Tiebout's now-classic paper (1956), fiscal economists have made significant progress in analyzing external exit. By "voting with their feet," individuals are able, in the limiting case, to ensure overall allocative efficiency in the supply of local public goods. Competitive governmental units are forced to supply these goods and services in preferred quantities and to "price" them, at least broadly, in line with relative-marginal evaluations. Elaborations of the analysis have largely involved departures from the limiting case assumptions through such elements as locational rents, the attenuation of ownership rights, interjurisdictional spillovers, absence of residual claimancy, nonhomogeneity in tastes, and nonconstant returns over membership sizes and public goods quantities.[1]

By contrast with these developments in analysis, almost no attention has been given to *internal exit*, which takes the form of secession by a coalition of people from an existing political unit along with the establishment of a new political unit that will then provide public goods to those who defect from the original unit. The neglect of internal exit may be due, in part, to the implicit presupposition that secession is legally, constitutionally impermissible, and, in part, to the unexamined assumption that secession is prohibitively costly due either to the locational interdependence among people in a polity or to the difficulties of forming coalitions among potential members of any seceding group. But if other margins of adjustment are costly or foreclosed, and if organizational difficulties are overcome, internal exit becomes a viable mode of response.

Since we do observe local governments that are clearly competitive and at least some competitiveness even among nations, the external-exit model both seems to be, and is, more "realistic" than its internal-exit counterpart. However, secession in various forms does occur. An example is incorporation in which a subset of citizens of an existing jurisdiction, say a county, set up their own policy, and provide and finance by themselves many of the public goods provided by the county. More dramatic examples of secession are threats and declarations of independence from existing national governments. If, in fact, successful secession is not often observed to occur in practice, the existence of pressures for internal exit may exert effects on the behavior of governments. That is to say, internal exit may be a "road not traveled," save under exceptional circumstances. The existence of such an alternative opportunity, along with its characteristic features, must, however, affect the attitudinal stance of people in their acquiescence in and/or criticism of political decisions beyond their individual control. For example, taxation beyond the limits defined by a plausibly estimated internal-exit option may erode the moral basis necessary for essentially voluntary tax compliance. Indirectly, a model of internal exit may be helpful in deriving testable implications relative to the growth of tax evasion-avoidance or, conversely, to the pressures for tax reform effort.

*Center for Study of Public Choice, George Mason University, Fairfax, VA 22030, and Department of Economics, Arizona State University, Tempe, AZ 85287, respectively. We would like to thank referees for some helpful suggestions.

[1]The paper by J. Vernon Henderson (1985) is the latest in a long series. See, especially, James Buchanan and Charles Goetz, 1972; F. Flatters, J. Vernon Henderson, and Peter Mieszkowski, 1974; and Eitan Berglas, 1976.

This paper introduces such a model. We shall assume throughout that people possess legal-constitutional rights of secession, which may or may not prove costly to exercise. It is immediately evident that any such liberty imposes constraints on the potentially exploitative behavior of those in the dominating or ruling political coalition, which, for reasons that will be apparent, we call the *sharing coalition*. The ability of members of this coalition to extract fiscal surplus is potentially restricted in ways that are analogous to those present under the external-exit prospect.

A broad interpretation of the sharing coalition includes all groups that are successful in obtaining net transfers from the government. These groups, by actively participating in the political process, may be able to redistribute wealth from other unorganized or less effective groups in the polity. The people or groups remaining outside the sharing coalition might represent the politically ineffective, unrepresented, or rationally nonparticipating segment of the population. But such a group also presents a potential for secession—literally or figuratively, such as evading taxes or withholding moral support for the institutions of governance.

The government provides public goods here assumed to be inherently monopolistic. The minimal rate of tax, therefore, for a polity of any size, is the rate that generates just sufficient revenue to finance the production of the public goods. But the monopolist provider (an individual or a coalition) will also try, to the extent that is possible, to use the taxing power to transfer revenues to itself. We proceed under the highly stylized assumptions that taxes must be levied proportionately on incomes of *all* members of the polity, whether in or outside the sharing coalition, that any available fiscal surplus must be shared *equally* among people who are members of the sharing coalition, and that no person outside the sharing coalition receive any fiscal transfers.[2] Given this stylized setting, the critical assumption that places limits on the amount of fiscal surplus is the liberty of secession. Individuals in any size group may secede and, without cost, form a new political unit. This unit, once it has seceded, must provide its own public goods.

The effectiveness of potential secession, or conversely, the extent of possible surplus extraction, depends on such parameters as the costs and the publicness of the services that are collectively financed, the form of the production function in the market sector, the differentiation of people in economic characteristics, and the overall size of the polity's membership. The extent of possible surplus extraction is also affected by the threat of change in the *composition* of the sharing coalition in government, by either political competition or more violent means. Certainly, seeking change in the composition of the sharing coalition is a margin on which individual response may take place along with or in place of internal exit. In our analysis, we concentrate solely on the secession option.

We shall see how the tax rate, total transfers, the gain from entry into the sharing coalition, and the effect of entry on existing sharers change as the sharing coalition increases in size. We shall show that the optimal-sized coalition from a sharer's viewpoint is not the smallest possible coalition due to scale economies in the production of private and public goods. In fact, over a wide range of sizes of the sharing coalition, entry into the government may be encouraged. We shall also concentrate on the possible conflicts that arise between sharers and nonsharers regarding entry into the sharing coalition and entry into the original polity. How entry is achieved or prevented, how the composition of the sharing coalition is determined, and how nonsharers choose between secession and other modes of response such as external exit, voice, or revolution will not be addressed here.

[2] In general terms, such asymmetry between the taxing and spending sides is historically descriptive of the constitutionally constrained U.S. fiscal structure. The uniformity clause severely restricts discrimination in taxation; no such limit is imposed on the spending or transfer side of budget.

I. The General Model

Government has a necessary function; it must provide "order," a nonexclusive, lumpy and costly good. Without "order" there is no private product. The cost of providing the required amount of the public good to a community of K people is $f(K)$, where $f'(K) \geq 0$. Each person in a community of K has a private product, or income, of $g(K)$. We assume $Kg(K) > f(K)$. The original polity consists of N identical individuals, M of them belonging to the government, or sharing coalition. The remaining $S = N - M$ individuals form the set of potential seceders.

Total fiscal surplus, or transfers, T is the difference between the tax revenue obtained by levying a nondiscriminatory tax rate, t, on private incomes, minus the total cost of the public good. Each member of the sharing coalition has a post-tax net income B equal to his post-tax private income $P = g(K)(1-t)$ plus an equal share T/M of the fiscal surplus.[3] A nonsharer's post-tax net income is simply P.

An equilibrium tax rate is one which given M and N maximizes the post-tax net income of the sharers without inducing secession.

In a community of K people, the minimal, nonexploitative tax rate is $t_0(K) = f(K)/Kg(K)$, which generates just enough revenue to finance the public good. Since the S nonsharers on their own in a new polity realize a post-tax income of $g(S)(1-t_0(S))$, the maximum tax rate a sharing coalition of size M in a polity of size N can levy without inducing secession is $t^*(M,N) = t_0(S) = f(S)/Sg(S)$.[4] Since B is an increasing function of t, the tax rate $t^*(M,N)$ is an equilibrium rate.

[3] To avoid problems of indifference, we shall assume that nonsharers are given a tiny amount of the fiscal surplus, $T - f(N)$.

[4] This tax rate assumes the seceders will employ the nonexploitative tax rate to finance the public good in the new polity. It might be argued that even in the new polity, the government will again behave in an exploitative manner. If so, it turns out that as long as each potential seceder sees himself having an equal probability M/S of belonging to the new sharing coalition, the maximal tax rate in the original polity will equal $f(S)/Sg(S)$. See James Buchanan and Roger Faith (1986).

Substituting t^* into T, P, and B, we obtain

$$(1)\quad T = t^*Ng(N) - f(N) = \left[\frac{f(S)}{Sg(S)} - \frac{f(N)}{Ng(N)}\right]Ng(N),$$

$$(2)\quad P = g(N)(1-t^*) = g(N) - \frac{g(N)f(S)}{Sg(S)},$$

$$(3)\quad B = \frac{T}{M} + P = [t^*Ng(N) - f(N)]/M + g(N)(1-t^*).$$

If $f(S)/Sg(S) > f(N)/Ng(N)$ for all $S < N$, that is, if the ratio of total cost to total product is lower for a polity of size N than for any smaller-size polity, transfers are positive.

Differentiating (1) through (3) with respect to M yields

$$(4)\quad T_M = Ng(N)t_M^*,$$

$$(5)\quad P_M = -g(N)t_M^*,$$

$$(6)\quad B_M = \frac{1}{M^2}[t_M^*SMg(N) - T].$$

The sign of t_M^* is the same as the sign of T_M and opposite the sign of P_M. Thus, if the maximum tax rate increases with the size of the sharing coalition, transfers increase and the post-tax position of nonsharers decreases. Because T is positive, a necessary condition for $B_M > 0$ is $t_M^* > 0$. However, the positivity of t_M^* does not imply the positivity of B_M. The reason is that as M increases and transfers increase, the increased transfers are divided over a larger group, and the tax paid by each sharer also rises, canceling some of the gain from the increase in transfers.

Differentiating t^* with respect to M reveals that

$$(7)\quad t^*_M \gtrless 0 \rightleftarrows \frac{Sg'(S)+g(S)}{g(S)} \gtrless \frac{Sf'(S)}{f(S)}.$$

The maximum tax rate increases in M, if and only if the elasticity of total product $Sg(S)$ with respect to community size is greater than the size elasticity of the total public good cost $f(S)$. A larger sharing coalition means a smaller population, a lower total product and tax base, and a lower total cost of the public good in the secessionist polity. If the size elasticity of product is greater than the size elasticity of cost, the tax base falls more than cost as the new polity shrinks in size and the tax rate necessary to finance the public good increases.

Using t^* and t^*_M in (6), rearranging and multiplying by S, we find

$$(8)\quad B_M \gtrless 0$$
$$\rightleftarrows \left[\frac{Sg'(S)+g(S)}{g(S)} - \frac{Sf'(S)}{f(S)}\right]$$
$$\gtrless \frac{N}{M}\left[1-\frac{f(N)}{f(S)}\frac{g(S)S}{g(N)N}\right].$$

The first bracketed term is the difference between the size elasticities of private product and public good cost. By (7), this difference is positive if $t^*_M > 0$. If $f(S)/Sg(S) > f(N)/Ng(N)$ for all S, then the second bracketed term is also positive, and B_M may be positive or negative.

The difference in the conditions for positivity of t^*_M and B_M implies a potential conflict between existing and potential members of the sharing coalition. Nonsharers will seek entry into a sharing coalition of size M if the net gain G to the potential entrant is positive,

$$(9)\qquad G = B(M) + B_M(M) - P(M)$$
$$\cong B_M + \frac{T}{M} > 0.$$

Using (4) and (6) in (9) shows that $G > 0$ if $t^*_M > 0$. The effect of entry on current sharers is B_M, which as just shown may be negative even if $t^*_M > 0$. If $B_M > 0$, entry into the sharing coalition will be encouraged by current sharers. If $B_M < 0$, current sharers will resist attempts by current nonsharers to join the sharing coalition. If entry beyond the point where $B_M = 0$ is precluded by barriers to entry not discussed here, those denied entry will not secede. Equilibrium implies unequal treatment of fiscal equals. If there are no barriers to entry and $t^*_M > 0$ for all $M < N$, then the size of the sharing coalition will be N, with an equilibrium tax rate $t^* = t_0(N) = f(N)/Ng(N)$. There is no exploitation.

It is clear from the discussion that the amount of fiscal exploitation and the tendency for the government to get larger depend on the behavior of the product and cost functions. Note, however, that if the cost of the public good does not increase less than proportionately with private production, there is no efficiency-based argument for collective or public provision of any good stemming from nonrivalry in usage. Only technological nonexclusiveness might then justify collective provision.

A. *Some Illustrative Cases*

Assume that the average cost over individuals, $f(K)/K$, falls in K, and average product $g(K)$ is constant and equal to $\bar{g}$.[5] Declining average cost is standard in public-goods theory which postulates the existence of some degree of nonrivalry in consumption, hence, increasing returns over group size, at least over some ranges. Constant average product may be interpreted as an absence of agglomeration economies.

[5]If we assume that average public goods cost rises with polity size, in the absence of any locational rents, the polity would break apart into smaller polities to avoid the diseconomies of scale. This would occur independent of the taxing proclivity of the government. If average cost were U-shaped, then we expect the polity would again break into several smaller communities, perhaps until minimum average cost is reached. At that size, any further size reductions would result in an increase in average cost and our model would apply.

Equations (1), (2), and (3) now become

$$(10)\quad T = N\left[\frac{f(S)}{S} - \frac{f(N)}{N}\right],$$

$$(11)\quad P = \bar{g} - \frac{f(S)}{S},$$

$$(12)\quad B = \frac{T}{M} + P = \frac{f(S) - f(N)}{M} + \bar{g}.$$

Since $f(N)/N < f(S)/S$, then $T > 0$ and $B > P$. Any secessionist group faces a higher average cost of producing the public good. The sharing coalition, in essence, taxes the nonsharers for the benefit of living in a larger polity—the reduction in the average cost of the public good. Using (7) and (8), gives

$$(13)\quad t_M^* \underset{<}{\overset{>}{-}} 0 \rightleftarrows \frac{f(S)}{S} - f'(S) \underset{<}{\overset{>}{-}} 0,$$

$$(14)\quad B_M \underset{<}{\overset{>}{-}} 0 \rightleftarrows \frac{f(S)}{S} - f'(S) \underset{<}{\overset{>}{-}} \frac{N}{M}\left[\frac{f(S)}{S} - \frac{f(N)}{N}\right].$$

Expression (13) says that the maximum tax rate increasing in M is equivalent to declining average cost. Expression (14) says that for entry to go unchallenged, average cost must not only decline with polity size, but do so at a sufficiently high rate.

If the government-provided good ("order" in our example) is a pure public good, then $f(K) = F$, and equations (10), (11), and (12) become

$$T = \frac{M}{N-M}F, \quad P = \bar{g} - \frac{F}{N-M}, \quad B = \bar{g}.$$

The value of B is independent of coalition size, therefore entry is costless to the sharing coalition until $M = N - 1$. Because *total* cost of the public good is constant, the sharing coalition taxes away from the nonsharers the benefits of dividing cost over a larger group size.

Assume now that average public good cost is constant over polity size but that larger polities have greater per capita private product. If there are agglomeration economies, larger polities will generate proportionally more total product. If part of the original polity secedes, the secessionists face a lower average product (on the assumption that inter-polity trade does not emerge to capture gains of interdependence) and a higher tax rate.

The basic relationships given $f(K)/K = \bar{f}$, are

$$(15)\quad T = \bar{f}N\frac{g(N) - g(S)}{g(S)},$$

$$(16)\quad P = g(N) - \frac{\bar{f}g(N)}{g(S)}, \text{ and}$$

$$(17)\quad B = \frac{\bar{f}N}{M}\frac{g(N) - g(S)}{g(S)} + g(N) - \frac{\bar{f}g(N)}{g(S)}.$$

Since $g(N) \geq g(S)$, then $B \geq g(N)$ and $P \leq g(N)$. Members of the sharing coalition can improve their lot over and above the "free" consumption of the public good. Even though average public good cost does not change, the smaller private productive capacity of the seceding group relative to the original polity implies that the opportunity costs of leaving are higher than before. This difference is recognized and taken advantage of by the sharing coalition.

Expressions (7) and (8) become

$$(18)\quad t_M^* \underset{<}{\overset{>}{-}} 0 \rightleftarrows \frac{Sg'(S)}{g(S)} \underset{<}{\overset{>}{-}} 0,$$

$$(19)\quad B_M \underset{<}{\overset{>}{-}} 0 \rightleftarrows [Sg'(S) + g(S)] - g(S) \underset{<}{\overset{>}{-}} \frac{N}{M}[g(N) - g(S)]\frac{g(S)}{g(N)}.$$

Note that increasing average product is necessary but not sufficient for entry to not harm the existing sharing coalition.

B. *Price-Elastic Public Good Demand*

We have assumed that the government-produced good, order, is consumable in only one quantity. Either there is order or there is not. Alternatively, we can say that the demand for order is perfectly inelastic at the required quantity. Suppose, instead, that the government provides goods such as schools and roads, which exhibit nonzero price elasticity. The nonsharers can provide the public good at the tax rate $t_0 = f(Q^S(t_0))/Sg(S)$, where $Q^S(t_0)$ is the quantity demanded by the nonsharers at the tax rate t_0. From our previous discussions we know that the equilibrium tax rate in the original polity t^* equals $t_0(S)$.

The quantity produced in the original polity, we assume, is determined by the sharing coalition and depends on the *effective*, post-transfer tax rate paid by the sharers defined as the equilibrium rate t^* minus the transfers per unit of income $T/Mg(N)$. Given $t^* = t_0$, the effective tax rate is

$$t_e = \frac{1}{M}\left[\frac{f(Q^M(t_e))}{g(N)} - \frac{f(Q^S(t_0))}{g(S)}\right], \tag{20}$$

where $Q^M(t_e)$ is the quantity demanded by the sharing coalition at the effective tax rate t_e.[6]

If cost is an increasing function of output, then $f(Q^S(t_0))$ is less than $f(Q^M(t_e))$, since t_0 is greater than t_e and $Q^S(t_0)$ is less than $Q^M(t_e)$. Thus, the more price elastic the demand for the public good, (evaluated at Q^S), the lower the equilibrium tax rate and the more effectively the liberty of secession constrains the government's ability to transfer.

[6]This assumes no congestion effect. A combination of congestion and price effects may act to cancel or reinforce each other in determining the magnitude of the equilibrium tax rate in the N-sized polity (see Buchanan and Faith, 1986).

II. Two-Class Model

The analysis to this point has proceeded under the assumption that individuals are identical in terms of their income levels. The *identity* of those making up the sharing coalition does not emerge as a relevant consideration. We shall now relax this assumption.

Assume a polity of size N consists of two internally homogeneous groups of individuals, high-income (H) and low-income (L), so that $N_H + N_L = N$. Total product in the polity is $N_H g_H + N_L g_L = W$. The equilibrium tax rate is

$$t^* = \frac{f(S)}{N_{HS}g_H + N_{LS}g_L} = \frac{f(S)}{W_S}, \tag{21}$$

where N_{ij} is the number of people of a particular income level in a particular group (for example, N_{LS} is the number of low-income people in the nonsharing group).

Our basic relationships, using (21) are

$$T = f(S)\frac{W}{W_S} - f(N), \tag{22}$$

$$P_i = g_i\left(1 - \frac{f(S)}{W_S}\right), \quad i = H, L; \quad \text{and}$$

$$B_i = \frac{1}{M}\left[\frac{f(S)W}{W_S} - f(N)\right] + g_i\left(1 - \frac{f(S)}{W_S}\right), \quad i = H, L. \tag{24}$$

Equation (21) implies that sharing groups with a greater proportion of low-income people are taxed more heavily than nonsharing groups with a greater fraction of high-income people. Now consider the difference between moving one high-income person from the nonsharing group to the sharing group, and moving one low-income person from the nonsharing to the sharing group. The differential effect on any original mem-

ber of the sharing coalition with income g_i is

$$(25)\qquad (g_H - g_L)\frac{f(S)}{W_S^2}\left(\frac{W}{M} - g_i\right),\qquad i = H, L.$$

For a low-income member of the sharing coalition, (25) is always positive; while for a high-income member, (25) is positive (negative), if $((W/M) - g_H)$ is positive (negative).[7] That is, it typically benefits the members of the sharing coalition *more* to add a high-income person rather than a low-income person to the sharing coalition. Equation (25) implies that the greater the income difference between classes $(g_H - g_L)$, the greater the incentive to bring high-income people into the sharing coalition.

This result has a powerful implication. It suggests that if the sharing coalition can control entry the government will tend to be dominated by people with relatively high income. Indeed, even if the original members of the government are low-income persons, they will prefer that a high-income person join the sharing group rather than another low-income person.

A. *The Effect of Polity Size and Immigration*

We now permit polity size to change through immigration. For ease in interpretation, we shall assume no congestion—$f(K) = F$, all K. Assuming new citizens enter the polity as nonsharers, the marginal effects on sharers' and nonsharers' net incomes are

$$(26)\qquad \frac{\partial B_i}{\partial N_{js}} = \frac{F g_j}{W_S^2}\left(\frac{W_S - W}{M} + g_i\right),\qquad i, j = L, H$$

$$(27)\qquad \frac{\partial P_i}{\partial N_{js}} = g_i g_j \frac{F}{W_S^2} > 0.\qquad i, j = L, H.$$

Because $g_H \geq (W - W_S)/M$ and $g_L \leq (W - W_S)/M$, equation (26) is nonpositive for low-income sharers and nonnegative for high-income sharers regardless of the income of the newcomer j. Adding another person to the polity as an outsider cannot harm and may benefit high-income members of the sharing coalition and cannot benefit and may harm low-income members of the sharing coalition.[8] Equation (27), however, is always positive, indicating that current nonsharers do better if their group grows since it lowers the equilibrium tax rate. Note also that $\partial P_i/\partial N_{LS} < \partial P_i/\partial N_{HS}$. Any nonsharer, regardless of income, prefers high rather than low-income individuals to enter the polity. Thus, conflict over immigration policy—who, if anyone at all, shall be allowed to *enter the polity*—will arise when incomes differ in the original polity.

B. *Income Elastic Public Good Demand*

In Section I, Part B, we discussed the effect of nonzero price elasticity for public goods on the model. In the two-class model, similar effects are produced if demand is income elastic. Specifically, we shall assume that the demand for the public good depends

[7]Bringing a high-income person versus a low-income person into the sharing coalition generates differentially higher tax rates. Intuitively, it is possible that if a coalition has already attained sufficient size and if a member of the coalition is sufficiently rich relative to low-income individuals, the higher tax rate costs him more in taxes than he gains in increased transfers. The sufficient condition for this possibility, $g_H > W/M$, can be rewritten as $(g_H/g_L) > N_L/(N_{LM} - N_{HS})$. The greater the high-low income ratio, the greater the likelihood that a high-income member of the sharing coalition will prefer a *low-income* rather than a *high-income* entrant.

[8]The entry of a new nonsharing person to the polity increases the aggregate tax base, but reduces the maximally exploitative tax rate. The low-income member of the sharing coalition secures some part of his transfer from high-income members of the coalition. Since the maximum tax rate falls, the within-coalition part of the transfer falls, hence, benefiting the high-income member and harming the low-income member of the sharing coalition.

on average income in the polity, or $Q^K(\overline{W}_K)$; $\partial Q^K/\partial \overline{W}_K > 0$, where $\overline{W}_K = W_K/K$. Assuming no congestion or price effects, the equilibrium tax rate is $t^* = f(Q^S(\overline{W}_S))/W_S$, and

$$(28) \qquad B_i = \frac{1}{M}\left[\frac{f\big(Q^S(\overline{W}_S)\big)W}{W_S} - f\big(Q^M(\overline{W}_M)\big)\right] + g_i\left(1 - \frac{f\big(Q^S(\overline{W}_S)\big)}{W_S}\right), \quad i = L, H,$$

$$(29) \qquad P_i = g_i\left(1 - \frac{f\big(Q^S(\overline{W}_S)\big)}{W_S}\right) \quad i = L, H.$$[9]

Assuming $\overline{W}_M > \overline{W}_S$, the nonsharing group imposes a stricter constraint on the sharing coalition than in the income-inelastic case, since nonsharers demand less of the public good and therefore face lower total costs of producing the public good. The greater the income elasticity, the greater the reduction in the equilibrium tax rate.

Other effects are generated also. First, the sharing coalition's incentive to favor rich over poor entrants into the coalition is reduced, since the higher the average income in the nonsharing group, the greater the nonsharers' demand for the public good. This means a higher equilibrium tax rate and greater transfers. Next, if average private product is subject to agglomeration economies, a positive income elasticity partially offsets the effects of lost economies of scale in the secessionist polity. Third, in an external-exit model, individuals of like income tend to group together because their public good demands are similar and group welfare is enhanced. In the internal-exit model, the potential for transfers may outweigh the efficiency gains from attaining unanimous agreement over the quantity of public goods. This suggests there are forces which keep polities of mixed income intact.

[9]The effects of nonzero price elasticity, income elasticity, and congestion can be combined to generate a more general set of relationships. See Buchanan and Faith (1986) for details.

III. Implications and Extensions

We have subtitled this paper "*Toward* a Theory of Internal Exit," and we have stressed the restrictiveness of the assumptions within which our basic models have been presented.

Despite the restrictiveness of the models, to us the implications derived are both interesting and potentially relevant. First, in the absence of elements not considered here, the size of the governing coalition tends to increase. There is an asymmetry which ensures that over a range of sizes the costs of new entrants, to members of any initial ruling coalition, are less than the benefits to those who are successful in securing entry. As the size of the sharing coalition increases, absolute tax rates increase, the total revenue side of the budget increases, and the relative size of transfers in the budget increases. Appealing to the conditions for unchallenged entry, the government will grow where per capita public goods cost is declining relative to average private product. If we may characterize mature economies as ones in which average private product growth slows, as population grows, and the average cost of order increases, due, perhaps, to increased congestion, we may expect to find greater resistance to increasing the sharing coalition from those within the coalition. In young, vigorous economies, the opposite may be expected.

A second interesting result emerges when we allow for differences among the economic positions of those who might be members of the ruling coalition. The rich are favored over the poor as potential entrants into the sharing group, quite independently of the identity of those who might initially hold the power of governance. Outside the coalition, the rich can, under our assumptions, more readily set up seceding polities. Hence, the

threat that they will do so must reduce the maximal tax that may be imposed. Further, once the rich are inside the sharing group, as both transfer recipients and as taxpayers generally, those within the group who are poor can gain. Once all of the rich are within the sharing coalition, the additional entry of members who are poor will tend to be opposed. Those who are poor remain outside the sharing coalition and, because they remain poor, they cannot readily secede. They either remain subject to maximal fiscal exploitation or possibly resort to extreme measures such as revolution. This seems to be broadly descriptive of modern politics despite the extreme restrictiveness of our model. The transfer state of modern political reality is not the transfer state as idealized by the egalitarian philosophers.

Third, the model implies that expansion policy will be viewed differently by the sharers and nonsharers. An example is annexation. Whether one community votes to absorb another community depends, in part, on the fiscal effects of the absorbing new citizens on existing citizens, some of whom may be classified as net beneficiaries (sharers) and some as net tax contributors (nonsharers) of local public services.

Our model also applies to any organization which provides collective benefits to its members, with the authority to tax its members, and competing organizations are not cheaply available. Social clubs, religious groups, home-owner associations, and labor unions are potential examples.

Implicitly, Tiebout's model of local government behavior offered a positive theory of the limits of taxation by governmental units in competition, one with another. There has been no comparable effort to work out, either implicitly or explicitly, a positive theory of absolute limits to taxation in a polity where external exit is not available to members. Realistically, of course, secession is less likely to be observed than direct overturn of governments, whether through democratic or nondemocratic means. The secession models presented in this paper would seem to provide a useful beginning stage for the more complex analysis that would be required to examine how the prospects for removal from authority might exert limits on the taxing proclivity of government.

REFERENCES

Berglas, Eitan, "Distribution of Tastes and Skills and the Provision of Local Public Goods," *Journal of Public Economics*, November 1976, *6*, 409–23.

Buchanan, James, and Faith, Roger, "Secession and the Sharing of Surplus," Working Paper, Arizona State University, October 1986.

______ **and Goetz, Charles,** "Efficiency Limits of Fiscal Mobility: An Assessment of the Tiebout Model," *Journal of Public Economics*, April 1972, *1*, 25–43.

Flatters, F., Henderson, J. Vernon and Mieszkowski, Peter, "Public Goods, Efficiency, and Regional Fiscal Equalization," *Journal of Public Economics*, May 1974, *3*, 99–112.

Henderson, J. Vernon, "The Tiebout Model: Bring Back the Entrepreneurs," *Journal of Political Economy*, April 1985, *83*, 248–64.

Tiebout, Charles, "A Pure Theory of Local Expenditures," *Journal of Political Economy*, October 1956, *64*, 416–24.

Honesty in a Model of Strategic Information Transmission

By CAROLYN PITCHIK AND ANDREW SCHOTTER*

We consider the following simple model of consumer fraud. Consumers need one of two possible repairs—an expensive repair (denoted E) or an inexpensive repair (denoted I). The exogenous probability that a consumer needs the more costly remedy is r. Each consumer is assumed to know r. An expert can observe with certainty which of the two repairs is needed and can offer to sell either to the consumer. (In a simple extension, we allow for uncertainty with respect to the expert's observation.) We assume that selling the expensive remedy to a consumer who needs only the inexpensive one is more profitable than selling the inexpensive one. Thus, $\Pi(E|I) > \Pi(I|I)$, where Π denotes profit and $|$ denotes "conditional on needing." The profit functions are increasing in the price of the respective repairs. (For convenience, we suppress the dependence of Π on prices in our notation.) We also assume that a consumer prefers to obtain the appropriate repair to the inappropriate one. Thus, $u(E|E) > u(I|E)$ and $u(I|I) > u(E|I)$, where u is the consumer's payoff. Lastly, we assume that the functions $u(E|E)$ and $u(E|I)$ are decreasing in the price of the expensive repair and that $u(I|E)$ and $u(I|I)$ are decreasing in the price of the inexpensive repair. We also assume that $u(I|E)$ is decreasing in the price of the expensive repair since the consumer ultimately buys the expensive repair in this case. (See below for further assumptions regarding the repairs.)

We are interested in the equilibrium levels of honesty in this model. How honest is the expert at the equilibrium? Do consumers always follow the expert's advice? Does the level of honesty increase as the interests of the agents become more similar (in a sense to be defined below)?

Our assumptions are similar to those in the abstract strategic information transmission models of Vincent Crawford and Joel Sobel (1982), and Jerry Green and Nancy Stokey (1980). We follow the former paper which asks whether the signals sent become less noisy as the agents' preferences become more similar. Despite the similarity of our model to Crawford-Sobel's, we arrive at some different conclusions with respect to the expert's honesty vis-à-vis the agents' closeness of interests. We explain this in detail in Section III. In addition we allow that the expert may be uncertain about the true state of the world. We also explicitly model the level of expert honesty, which extends Crawford-Sobel.

The outline of the paper is as follows. In Section I we provide the details of the model and show that a unique equilibrium exists. Comparative static results and a simple extension (incorporating the assumption of the incompetent expert) are given in Section II. We draw comparisons with Crawford-Sobel in Section III. In Section IV we conclude with a brief summary.

I. The Model and the Solution

The model works as follows. First, the expert, knowing the true repair needed, announces the sale of one of the two available repairs to the consumer. The consumer can choose either to obtain this repair or to disregard the advice. For simplicity, we as-

*Departments of Economics, University of Toronto, Toronto, Ontario, Canada M5S 1A1; and New York University, New York NY 10003, respectively. We are especially grateful to Clive Bull and Martin Osborne for their many helpful comments and constructive criticisms. We also thank Jess Benhabib, Ariel Rubinstein, and Bernard Wasow for their comments on an earlier draft of this paper. We thank two anonymous referees for comments leading to this final revision. We acknowledge the National Science Foundation grant no. SES-8207765, the Office of Naval Research contract no. N0014-78-0598, and the Social Sciences and Humanities Council of Canada for their partial support.

sume that whenever the expert's advice is ignored, the consumer obtains a remedy from someone else. (For example, in the automobile repair industry, do-it-yourself repairs are possible. In the medical industry, second opinions and alternative treatments are available.)

Finally, we make three additional assumptions concerning the repairs. We assume that the consumer can detect only whether the problem still exists. If the problem no longer exists, the type of repair actually performed is unknown. Thus, the repair is a credence good (see Michael Darby and Edi Karni, 1973). In addition, we assume that the expensive remedy repairs either problem, while the inexpensive one is only good for the inexpensive problem. Suppose that an expert advises a consumer to obtain an inappropriate inexpensive repair. If the consumer acts on this advice, the consumer's problem remains unrepaired. However, then it is clear that the expert lied to the consumer. These assumptions thus make it reasonable to suppose that a liability rule is in effect concerning the sale of an inexpensive repair when an expensive one is necessary. The rule makes the expert liable for advice, which results in the consumer's obtaining an inappropriate inexpensive repair. (For example, a liability rule might require the expert to refund any expenses that the consumer incurs in such an event.) Such a liability rule prevents the expert from profiting by advising consumers to obtain the less expensive remedy when the more expensive one is needed.

The model is described by the strategic game in Table 1, once duplicated and dominated strategies have been deleted. The strategies of the consumer and the expert are as follows. The first character in the consumer's strategy indicates acceptance (A) or rejection (R) of advice to obtain the expensive repair; the second character, acceptance or rejection of inexpensive repair advice. The first character in the expert's strategy indicates telling the truth (T) or not (N), if the true repair needed is the expensive one; the second character, telling the truth or not, if the inexpensive one is needed. (Because of the liability rule, AA and RA dominate AR and RR, respectively, and TT and TN dominate NT and NN, respectively. Thus, the strategic form is as described in Table 1, once dominated and duplicated strategies have been eliminated.) The assumptions on the payoffs made earlier are equivalent to the following:[1] $E_2 > E_1$, $E_3 > E_4$, $C_1 > C_3$. We make the additional assumption that $C_4 > C_2$ in order to create the underlying tension. Otherwise, the expert always lies. (This last assumption is equivalent to an upper bound on r.)

TABLE 1—THE STRATEGIC GAME

		Expert	
		TT	TN
Consumer	AA	(C_1, E_1)	(C_2, E_2)
	RA	(C_3, E_3)	(C_4, E_4)

It is immediate that a unique equilibrium exists, and that this equilibrium is completely mixed. In the equilibrium the expert always tells the truth if the repair needed is the expensive one and tells the truth with probability $0 < \hat{p} < 1$, if the repair needed is the inexpensive one, where

$$(1)\quad \hat{p} = r[u(E|E) - u(I|E)] \div (1-r)[u(I|I) - u(E|I)].$$

The consumer always acts on the expert's advice to obtain the inexpensive repair, and acts on advice to obtain the expensive repair with probability $0 < \hat{q} < 1$, where

$$(2)\quad \hat{q} = \Pi(I|I)/\Pi(E|I).$$

In equilibrium the expert's payoff is $\hat{q}[r\Pi(E|E)+(1-r)\Pi(E|I)]$ and the consumer's is $ru(I|E)+(1-r)u(I|I)$.

[1]The payoffs are as follows: $C_1 = ru(E|E) + (1-r)u(I|I)$, $C_2 = ru(E|E) + (1-r)u(E|I)$, $C_3 = C_4 = u(I|E) + (1-r)u(I|I)$; $E_1 = r\Pi(E|E) + (1-r)\Pi(I|I)$, $E_2 = r\Pi(E|E) + (1-r)\Pi(E|I)$, $E_3 = (1-r)\Pi(I|I)$, $E_4 = 0$.

II. Comparative Statics

In this section we consider the effects of price and quality controls on expert honesty levels and utility levels. The quality we control is that of the product which is in need of repair.

A. *Quality Control*

The quality of the faulty product is indexed by r. Decreasing r lessens the probability that the repair needed is the expensive one. Suppose that r decreases. (This might be achieved by advertising preventive health measures in a medical context, or by regulating automobile product quality in an auto repair context.) To analyze this effect we need to know how both $\hat{p}$ and $\hat{q}$ depend on r. From equations (1) and (2) we see that $\hat{p}$ decreases in r and that $\hat{q}$ is independent of r. Thus, decreasing the probability that the repair needed is costly has the effect of increasing the expert's honesty. In effect, if it is less probable that a consumer needs the more expensive repair, then the expected cost of ignoring the expert's advice (when the expert advises an expensive repair) is lower. Consequently, the expert needs to be more honest than before to keep the consumer from always ignoring this advice. The consumer clearly prefers (i.e., has a higher payoff at) the equilibrium outcomes associated with a lower r, while the expert prefers the outcomes with a higher r (see fn. 1).

B. *Price Controls*

If the price of the expensive repair is decreased, then we assume that $\Pi(E|I)$ decreases, while the difference $\Pi(E|E)-\Pi(E|I)$ remains the same. In addition we assume that $u(E|E)$, $u(E|I)$ and $u(I|E)$ increase, while the difference $u(E|E)-u(E|I)$ remains the same. But then by equations (1) and (2), $\hat{p}$ and $\hat{q}$ both increase. Thus, decreasing the price of the more expensive repair increases both the probability that the expert is dishonest and that a consumer obtains the more costly remedy when advised to. The intuition is as follows. Decreasing the price of the more expensive repair decreases the expected cost of following the expert's advice to obtain the expensive repair. In effect, it allows the consumer to be less particular about choosing the expensive repair (i.e., $\hat{q}$ increases). This, in turn, allows the expert to exploit the consumer's loss of vigilance.

In addition, the lower the price of the expensive repair, the higher the consumer's expected payoff, and the lower the expert's expected profit at the equilibrium outcome (see fn. 1).

Lastly, we consider the effects of an increase in the price of the inexpensive repair. This increase is assumed to result in increases in $\Pi(I|I)$ and in $u(E|E)-u(I|E)$, and a decrease in $u(I|I)-u(E|I)$. Thus, using equations (1) and (2), we obtain higher values for both $\hat{q}$ and $\hat{p}$. In other words, increasing the price of the less costly remedy increases both the dishonesty of the expert and the probability that a consumer obtains the more expensive repair when advised to. The intuition is as follows. Increasing the price of the less costly remedy increases the expected cost of ignoring the expert's advice (when the expert advises an expensive repair). This makes the consumer less particular in listening to advice. This in turn allows the expert to be more dishonest.

With respect to equilibrium payoffs, the consumer is worse off, while the expert is better off at the equilibrium outcome associated with the higher price for the inexpensive repair (see fn. 1).

C. *Expert Competence*

Our model can easily incorporate the existence of an expert who does not always know with certainty the repair that the consumer needs. We extend the model simply by assuming that though an expert detects an expensive repair whenever one is needed, an inexpensive repair is detected with probability s. We can then consider the effects of an increase in expert competence.

We assume that the competence level s and the exogenous probability r are such that an expert's honest opinion is the best predictor of which repair is necessary. Otherwise, the assumptions remain the same.

As before $\hat{p}$ and $\hat{q}$ denote equilibrium values.

In this case,

$$[r(u(E|E)-u|I|E)) - \hat{p}=(1-r)(u|I|I)-u|E|I))(1-s)] \div [s(1-r)(u|I|I)-u|E|I))],$$

and $\hat{q}$ remains the same as before. We see that $\hat{p}$ increases in s and that $\hat{q}$ is independent of s. Thus, an increase in expert competence results in more expert dishonesty. The intuition is as follows. At any given dishonesty level of the expert, an increase in expert competence increases the validity of the expert's advice to obtain an expensive repair. The expert can then take advantage of the situation by becoming more dishonest without affecting the consumer's actions. The expert prefers the equilibrium outcome associated with a higher[2] s, while the consumer is indifferent between levels of[3] s. The earlier comparative static results remain valid under the assumption of an incompetent expert.

III. Comparison with Crawford and Sobel

In Crawford-Sobel, the signal sent to the receiver in equilibrium becomes more informative as the agents' preferences become more similar. In our model the expert does not become more honest in equilibrium as the agents become more similar. The difference lies in the meaning of "more similar."

In the continuous framework of Crawford-Sobel, agents become more similar as a particular parameter (b) of the model approaches zero. This parameter is an index of the distance between the agents' payoff functions. It is also an index of the distance between the receiver's actions which maximize each agent's payoff function in each state. As b becomes smaller, the signaler sends signals with finer partitions in the equilibrium associated with greater similarity of agents (i.e., with smaller b). Thus, as the agents become more similar, the signal sent in equilibrium is less noisy, and both agents are better off. In effect, the benefits from distortion are reduced as agents become more similar. However, as Crawford-Sobel note, a finer partition for the equilibrium signal is not operationally equivalent to the signaler's being more honest.

In our discrete model there are two states of the world and two possible actions the consumer may take. In one state (E) the preferences of the agents are the same. In this state the maximizer of each agents' payoff function with respect to the receiver's (i.e., the consumer's) actions is identical (and equal to E). In the other state (I) the preferences of the agents are opposed. The consumer's payoff-maximizer is I but the expert's is E. Given this discrete model, there is no change in parameters that can alter this divergence of preferences in state I. There is no parameter that plays a role analogous to that played by b in Crawford-Sobel. However, we can change the parameters in such a way that the agents become closer in different ways.

For example, the changes in prices that we consider all have the effect of making the payoff functions closer but leaving the payoff-maximizing actions the same. These price changes decrease the expert's potential gain from lying and the consumer's potential loss from choosing an inappropriate repair. In addition, the change in r that we consider alters the underlying distribution of states. An increase in r results in an increase in the probability of state (E), in which the agents' preferences coincide so that they are more likely *ex ante* to have congruent interests. In our model, if state I occurs, the signaler either sends I or sends E, each with positive probability, and thus is explicitly honest or dishonest. As the agents become more similar, in the sense just described, the signaler does not always become more honest (i.e., $\hat{p}$ does not always decrease). Nor are the agents both better off. Both of these results are in contrast to those in Crawford-Sobel.

[2]The expert's equilibrium expected payoff is $ru(E|E)+(1-r)[su(E|I)+(1-s)u(E|E)]$.

[3]The consumer's equilibrium expected payoff is $ru(I|E)+(1-r)u(I|I)$ (as before).

IV. Conclusion

In our discrete model of strategic information transmission, the expert explicitly chooses to be either honest or dishonest. Any change that decreases the consumer's cost of not heeding the expert's advice to obtain an expensive repair (i.e., decreasing the probability of needing the expensive repair, decreasing the price of either repair, or decreasing expert competence) results in increased consumer vigilance and expert honesty, and thus leaves the consumers better off and the experts worse off. Unlike the Crawford-Sobel model, agents are not both better off as they become more similar in the sense described above. We also consider changing the informational structure by allowing that the expert is not completely competent.

REFERENCES

Crawford, Vincent P. and Sobel, Joel, "Strategic Information Transmission," *Econometrica*, November 1982, *50*, 1431–451.

Darby, Michael and Karni, Edi, "Free Competition and the Optimal Amount of Fraud," *Journal of Law and Economics*, April 1973, *16*, 67–88.

Green, Jerry and Stokey, Nancy, "A Two-Person Game of Information Transmission," Discussion Paper No. 751, Harvard Institute of Economic Research, March 1980.

Operative Gift and Bequest Motives

By ANDREW B. ABEL*

In a pioneering paper, Robert Barro (1974) demonstrated that if consumers have operative altruistic bequest motives, then a reduction in lump-sum taxes, accompanied by the issue of an equal amount of government bonds, has no effect on the allocation of resources. Barro stressed that this result, which has come to be known as the Ricardian Equivalence Theorem, requires that the bequest motive be operative. In this context, the term "operative" means that equilibrium bequests are determined by tangency conditions rather than by corner solutions such as may arise from binding nonnegativity constraints. If the bequest motive is not operative, then the Ricardian equivalence result presented by Barro does not hold, and there are important effects associated with the government's choice between debt finance and taxes.

More recently, Willem Buiter (1979) and Jeffrey Carmichael (1982) have analyzed the altruistic gift motive in which consumers obtain utility from the utility of their parents, and thus may be motivated to give resources to their parents. Their analyses confirm Barro's claim (p. 1104) that if the gift motive is operative, then the Ricardian Equivalence Theorem holds. If the gift motive is not operative, then the Ricardian Equivalence Theorem fails to hold.

Because the Ricardian Equivalence Theorem depends on an operative motive for private intergenerational transfers, it is important to determine the conditions under which either transfer motive will be operative. Several papers have studied whether the bequest motive is operative in a variety of different models[1] but the literature does not contain an analysis of the conditions that determine whether the gift motive is operative. In this paper, I will study the conditions for an operative gift motive. However, rather than confine the analysis to a model in which consumers have only a gift motive, I will assume that individual consumers have two-sided transfer motives. That is, I will assume that individual consumers have both a gift motive and a bequest motive as in John Burbidge (1983), Buiter and Carmichael (1984), and Burbidge (1984).[2] In the steady-state equilibrium, the gift motive may be operative, the bequest motive may be operative, or neither motive may be operative. If either of the intergenerational transfer motives is operative, then the Ricardian Equivalence Theorem holds; however, if neither motive is operative, then changes in the timing of lump-sum taxes have important effects on the intertemporal and intergenerational allocation of resources.[3]

The major goal of this paper is to determine conditions under which each of the intergenerational transfer motives is operative if individual consumers have two-sided transfer motives. As a prerequisite to this analysis, I will discuss, in Section I, appropriate restrictions on the gift motive and the bequest motive. In Section II, I will discuss the restrictions on two-sided transfer motives implied by intergenerational consis-

*Department of Finance, Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6367. Previous versions of this paper were written while I was an Associate Professor at Harvard University. I thank B. Douglas Bernheim, Olivier Blanchard, Stanley Fischer, Miles Kimball, Lawrence Summers, and Philippe Weil for helpful discussions. Financial support from the Amoco Foundation Term Professorship in Finance, the National Science Foundation, and the Sloan Foundation is gratefully acknowledged.

[1] See Allan Drazen (1978), Weil (1987), Alex Cukierman (1986), Cukierman and Allan Meltzer (1986), Martin Feldstein (1976), and Abel (1986).

[2] Recently, Miles Kimball (1986) has extended the analysis in this paper to analyze the conditions under which there will be an operative bequest motive under two-sided altruism.

[3] As pointed out by Carmichael, in order for the Ricardian Equivalence Theorem to hold, the same transfer motive must be operative both before and after the change in fiscal policy.

tency. The specification of the motives for intergenerational transfers has important implications for a wide range of issues extending beyond the effects of fiscal policy, including the intergenerational transmission of inequality,[4] and for the behavior of financial markets, especially markets for life insurance and annuities.[5] In Section III I discuss the endogenous determination of equilibrium factor prices and then describe the steady-state equilibrium. The conditions under which one or the other of the transfer motives is operative are derived in Section IV. I present concluding remarks in Section V.

I. A Two-Sided Transfer Motive

In this section I present a two-sided transfer motive and discuss appropriate restrictions on the parameters of the transfer motive. Consider a representative consumer economy in which each consumer lives for two periods. A generation t consumer is born at the beginning of period t, consumes c_{1t} in period t at age 1 and consumes c_{2t+1} in period $t+1$ at age 2. Let $u_t = u(c_{1t}, c_{2t+1})$ be the utility that a generation t consumer obtains directly from his own consumption. Defining u_{1t} as $\partial u(c_{1t}, c_{2t+1})/\partial c_{1t}$ and u_{2t+1} as $\partial u(c_{1t}, c_{2t+1})/\partial c_{2t+1}$, assume that $u_{1t} > 0$, $u_{2t+1} > 0$ and that $u_{1t}(0,\cdot) = \infty = u_{2t+1}(\cdot,0)$. Also, assume that $u(\cdot,\cdot)$ is strictly concave and that c_{1t} and c_{2t+1} are normal goods.

In addition to obtaining utility directly from his own consumption, a generation t consumer obtains utility from the consumption of his parents and from the consumption of all of his descendants. In particular, I will use the Buiter-Carmichael (1984) generalization of the Burbidge (1983) two-sided utility function

$$v_t = u_t + \alpha u_{t-1} + \sum_{j=1}^{\infty} \beta^j u_{t+j}. \tag{1}$$

The parameter β measures the strength of the bequest motive and satisfies the restriction $0 \leq \beta < 1$. The assumption that β must be less than one is the standard assumption in the literature[6] and is necessary and sufficient for the transversality condition to hold in the steady state with constant per capita consumption. The nonnegative parameter α measures the strength of the gift motive. There is no compelling reason to restrict α to be less than one.[7,8] I will show in Section II that intergenerational consistency (defined below) places an upper bound on the admissible values of α, but depending on the value of β, this upper bound may be greater than, equal to, or less than one.

[4]See Andrew Abel (1985), Laurence Kotlikoff et al. (1984); and Nigel Tomes (1981).

[5]See, for example, Stanley Fischer (1973) and Benjamin Friedman and Mark Warshawsky (1984).

[6]See, for example, Buiter (1979), Buiter and Carmichael (1984), Carmichael (1982), Burbidge (1983, 1984), and Philippe Weil (1987).

[7]Buiter–Carmichael (1984) note that the specification of the gift motive as $v_t = u_t + \alpha v_{t-1}$ implies that $v_t = \sum_{j=0}^{\infty} \alpha^j u_{t-j}$. They argue that if $\alpha > 1$, then the utility v_t is unbounded as t approaches infinity. However, even if $\alpha \geq 1$, the maximization of (3) subject to the constraints on the generation t consumer is a well-defined maximization problem.

Alternatively, Buiter–Carmichael point out that if v_t is constant over time, then the "steady-state utility function" is $v(c_1, c_2) \equiv u(c_1, c_2)/[1-\alpha]$, where c_i is the steady-state consumption of consumers of age i. They observe that if $\alpha > 1$, then "the model has the peculiar characteristic that the steady-state utility function $v(\cdot)$ has the opposite properties to the consumption utility $u(\cdot)$; for example, if $u(\cdot)$ is positive and increasing in c_1 and c_2, $v(\cdot)$ is negative and *decreasing* in c_1 and c_2." (p. 763) However, the "steady-state utility function" $v(\cdot)$ is not a useful construct. Paul Samuelson (1968) showed that the steady-state capital stock is lower than the Golden Rule capital stock if consumption is allocated to maximize the weighted sum of utility of all generations, with declining weights on future generations (which is formally identical to the problem faced by consumers with a bequest motive in (2)). Maximization of the "steady-state utility function" led Buiter (1979) to conclude erroneously that if either the bequest motive or the gift motive is operative, then a competitive economy would attain the Golden Rule in the steady state and that "lump-sum taxation and debt policy will not affect the *steady-state* capital-labor ratio if there are both bequest and gift motives." (p. 425).

[8]In an interesting analysis of consumption and gift behavior under a specific assumption about expectations of future gifts, Hajime Hori and Jun Tsukamoto (1985) analyze the case in which $\alpha > 1$ as well as the case in which $\alpha < 1$.

The two-sided utility function in (1) nests both the one-sided altruistic bequest motive and the one-sided altruistic gift motive. The one-sided altruistic bequest motive is often specified recursively as

$$v_t = u_t + \beta v_{t+1}. \tag{2}$$

When $\alpha = 0$, the utility function in (1) is consistent with the recursively specified altruistic bequest motive in (2).[9]

The one-sided gift motive is often specified recursively as $v_t = u_t + \alpha v_{t-1}$, which can be rewritten as

$$v_t = u_t + \alpha u_{t-1} + \alpha^2 v_{t-2}. \tag{3}$$

From the point of view of the generation t consumer with the one-sided gift motive in (3), the utility of his grandparent, v_{t-2}, is fixed; maximization of the utility function in (3) is equivalent to maximization of the utility function in (1) when $\beta = 0$. Thus, the utility function in (1) nests the one-sided altruistic bequest motive and the one-sided altruistic gift motive.[10]

Before presenting the consumer's budget constraint, it is necessary to describe the demographic composition of dynastic families. Each consumer lives for two periods and has $n \geq 1$ children at the beginning of the second period of his life. This assumption follows the standard convention of ignoring the fact that it takes two people from different families to produce children.[11] In the model, each consumer has n children and has one parent.[12]

Let g_t be the gift given by a generation t consumer to his parent who is a generation $t-1$ consumer. This gift is made during period t which is the only period during which both generations are alive. Because the generation t consumer has one parent and n children, this consumer gives a gift of g_t in period t and receives gifts totaling ng_{t+1} in period $t+1$.

Let b_t be the bequest given by a generation t consumer to each of his n children (generation $t+1$ consumers) in period $t+1$. The generation t consumer receives a bequest b_{t-1} from his parent in period t. In addition to receiving the bequest b_{t-1} in period t, the generation t consumer inelastically supplies one unit of labor in period t and receives the real wage rate w_t in period t. The generation t consumer is retired in period $t+1$. Letting R_{t+1} be the gross rate of return on saving held from period t to period $t+1$, the budget constraint of a representative period t consumer is

$$\begin{aligned} [c_{1t} + g_t]R_{t+1} + c_{2t+1} + nb_t \\ = [w_t + b_{t-1}]R_{t+1} + ng_{t+1}. \end{aligned} \tag{4}$$

The left-hand side of (4) contains the generation t consumer's expenditure on his own consumption in the two periods of his life plus the expenditure on bequests to children and a gift to his parent. The right-hand side of (4) contains the three sources of the generation t consumer's resources: labor income, bequest received from his parent, and the gifts received from his children.

I use the standard Nash assumption that in choosing optimal values of consumption,

[9] Douglas Gale (1983) has pointed out that there is an infinity of infinite-horizon utility functions which are consistent with the recursive formulation in (2). By starting with equation (1) as the specification of preferences, I am explicitly choosing one solution, a practice which is followed, at least implicitly, in an overwhelming majority of the literature.

[10] The relation between the utility function in (1) and "two-sided altruism" is discussed in Miles Kimball (1986).

[11] Douglas Bernheim and Kyle Bagwell (1984) have recently provided a stimulating analysis of the implications of marriage and altruism for the efficacy of fiscal policy.

[12] This point has not been appreciated in the gift motive literature. In fairness to Carmichael, it must be noted that he seemed to be aware of this point and avoided its implications by treating the "descendents and forebearers as though there were only one of each; the descendent will be n times 'bigger,' and the forebearer n times 'smaller' than the individual." (1979, fn. 2). Subsequently, Buiter and Carmichael (1984, p. 763, fn. 2) recognized that each consumer has one, rather than $1/n$, (set of) parent(s). They use this observation to make an insightful comment on Burbidge's specification of the utility function, but they ignore this observation in deriving optimal individual behavior under the Nash assumption.

bequests, and gifts, the consumer takes as given the actions of all other members of his dynastic family. In particular, in choosing g_t, the generation t consumer takes as given the gifts given by his siblings to their common parent. The maximization problem of a representative generation t consumer is to maximize (1) subject to (4), the nonnegativity constraints[13] $g_t \geq 0$ and $b_t \geq 0$ and subject to the given values of the decisions of all other members of the dynastic family. Recalling that u_{1t} and u_{2t+1} are the derivatives of $u(c_{1t}, c_{2t+1})$ with respect to its first and second arguments, respectively, the first-order conditions are

$$u_{1t} = R_{t+1}u_{2t+1} \tag{5}$$

$$u_{1t} \geq \alpha u_{2t} \tag{6}$$

(holds with equality if $g_t > 0$)

$$u_{2t+1} \geq (\beta/n)u_{1t+1} \tag{7}$$

(holds with equality if $b_t > 0$)

Equation (5) characterizes the optimal intertemporal allocation of the consumer's own consumption over his lifetime. If the consumer reduces c_{1t} by one unit, he suffers a utility loss of u_{1t}. However, if this unit of the consumption good is saved, then c_{2t+1} can be increased by R_{t+1} units, which increases utility by $R_{t+1}u_{2t+1}$. At the optimum, the utility loss in period t is equated to the utility gain in period $t+1$, as indicated by (5).

Equation (6) characterizes the optimal gift g_t. In period t, the generation t consumer can reduce his own consumption by one unit, suffering a utility loss of u_{1t}, and can increase the gift g_t by one unit, increasing his parent's utility by u_{2t}. The increase in parent's utility raises the generation t consumer's utility by αu_{2t}. If the optimal gift is at an interior optimum ($g_t > 0$), then the utility loss (u_{1t}) from the reduction in c_{1t} will equal the utility gain (αu_{2t}) from the increased gift. If, at $g_t = 0$, the utility loss from reduced consumption exceeds the utility gain from an increased gift, then the consumer will not make a positive gift, and the nonnegativity constraint on the gift binds strictly. It is worth noting that if, for some unspecified reason, siblings jointly decide on the level of the gift to give to their common parent, or equivalently, if each consumer is assumed to have $1/n$ parents, then the first-order condition (6) must be amended to

$$u_{1t} \geq \alpha n u_{2t} \tag{6'}$$

(holds with equality if $g_t > 0$).

Equation (6′) corresponds to the first-order condition derived by Carmichael (1982) and is consistent with the conditions in Buiter and Carmichael (1984).

Equation (7) characterizes the optimal bequest b_t. The generation t consumer can reduce c_{2t+1} by one unit and increase the bequest to each child by $1/n$, which increases the utility of each child by $(1/n)u_{1t+1}$. If the bequest motive is operative ($b_t > 0$), then the utility loss from decreased consumption is equal to the utility gain from increasing the bequest. If the nonnegativity constraint binds strictly, then the inequality in (7) holds strictly.

II. Intergenerational Consistency Under a Two-Sided Motive

In this section I discuss the conditions under which the decisions of different generations within a family are "intergenerationally consistent." There are two aspects of intergenerational consistency. First, there is the notion of dynamic consistency introduced by Robert Strotz (1956). Strotz showed that for a particular formulation of the intertemporal utility function in which the discount factor between two periods depends only on the length of time between the two periods, and not on calendar time, the consumption plan will be dynamically inconsistent unless the discount factors are geo-

[13]The assumption that the marginal utility of consumption at each age becomes infinite as the level of consumption approaches zero implies that any nonnegativity constraints on consumption will not be binding.

metrically declining. In the context of the utility function in (1), it is important that the weights on u_{t+j} are geometrically declining for $j = 0, 1, 2, \ldots$. If these weights were not geometrically declining, then the consumption plan would suffer from dynamic inconsistency in Strotz's sense, if the bequest motive were operative.

The second notion of intergenerational consistency is that the first-order conditions of parents and their children should not contradict each other. More precisely, consider the first-order condition characterizing the optimal gift from a child to a parent at time t (equation (6)) and the first-order condition characterizing the optimal bequest from a parent to a child at time t (equation (7) with the time subscript decremented by 1). If both of these first-order conditions are to hold, then

$$(8) \qquad u_{1t} \geq \alpha u_{2t} \geq (\beta\alpha/n) u_{1t}.$$

Because u_{1t} is assumed to be positive, equation (8) implies that

$$(9) \qquad \beta\alpha \leq n.$$

Equation (9) along with the restrictions $0 \leq \beta < 1$ and $\alpha \geq 0$ describe the admissible values of the parameters α and β under the restriction that the two-sided transfer motive is intergenerationally consistent.

III. Competitive Factor Prices and Steady-State Equilibrium

In the previous sections I analyzed the behavior of an individual dynastic family taking as given the factor prices w_t and R_t. These factor prices, which are determined endogenously in competitive factor markets, depend on the productive technology. Let Y_t be gross output in period t. This output is homogenous and can either be consumed or used as capital in the following period. The level of output is determined by a neoclassical linearly homogeneous production function $Y_t = F(K_t, N_t)$, where K_t is the aggregate stock of capital and N_t is the number of young consumers who each supply one unit of labor. The production function $F(,)$ is a gross production function in the sense that the aggregate capital stock, K_{t+1}, is equal to output, Y_t, minus total consumption, $N_t c_{1t} + N_{t-1} c_{2t}$, in period t. The production function can be written in intensive form as $y = f(k)$, where y is the output-labor ratio, k is the capital-labor ratio, $f' > 0$ and $f'' < 0$.

In competitive factor markets, each factor is paid its marginal product

$$(10) \qquad R_t = R(k_t) \equiv f'(k_t)$$

$$(11) \qquad w_t = w(k_t) \equiv f(k_t) - k_t f'(k_t).$$

The steady state is characterized by constant values of consumption for both young consumers and old consumers. Therefore, u_{1t} and u_{2t} are each constant in the steady state. Equations (5)–(7) imply that in the steady state the interest rate R must satisfy the following condition

$$(12) \qquad \alpha \leq R \leq n/\beta.$$

If one of the transfer motives is operative, then the steady-state interest rate is at one of the boundaries in (12). In particular,

$$(13a) \qquad R = n/\beta \qquad \text{if } b > 0,$$

$$(13b) \qquad R = \alpha \qquad \text{if } g > 0.$$

Since β is restricted to be less than one, equation (13a) yields the well-known result that a steady state with operative bequests is undercapitalized relative to the Golden Rule (i.e., $R > n$). However, since α can be less than, greater than, or equal to n, equation (13b) implies that a steady state with an operative gift motive can be either overcapitalized, undercapitalized, or at the Golden Rule. This result is contrary to the result in Carmichael (1982) that a steady state with an operative gift motive is overcapitalized. Carmichael's overcapitalization result follows from his assumption that the gift parameter α must be less than one and from his implicit assumption that siblings jointly determine the gifts to their common parent according to (6′). Under this pair of assumptions, $R = n\alpha < n$ in the steady state with operative gifts.

Finally, observe from (13a, b) that if $\alpha\beta < n$, then either bequests or gifts must be equal to zero in the steady state. In the case with $\alpha\beta = n$, which is on the boundary of the admissible region of the parameter space, and which corresponds to Burbidge's specification,[14] it is possible for both gifts and bequests to be positive in the steady state. However, as shown below in Section IV, the direction of net intergenerational transfers will be determinate in this case. Also note that with $\alpha\beta = n$, the range of possible values for the steady-state interest rate in (12) is degenerate: the steady-state interest rate is equal to $n/\beta = \alpha$ regardless of the level of government debt that is serviced by lump-sum taxes. Finally, since at least one of the transfer motives is operative, the Ricardian Equivalence Theorem holds in this case, as argued by Burbidge.

IV. When Are the Transfer Motives Operative?

The neutrality of government debt requires that one of the transfer motives be operative both before and after the change in government debt, and furthermore, that the same motive be operative after the change as before the change. Since the Ricardian Equivalence Theorem rests on the existence of an operative transfer motive, the question of when one of the transfer motives will be operative takes on great importance. In this section, I extend Weil's (1987) analysis of the one-sided bequest motive in (2) to the case of the two-sided utility function in (1).

Recall that K_{t+1} is the total stock of capital at the beginning of period $t+1$. All of this capital is held by generation t consumers and, furthermore, this is the only asset held by these consumers. Therefore, letting s_t denote the saving of a representative generation t consumer, it follows that $K_{t+1} = N_t s_t$, which can be written as

$$nk_{t+1} = s_t. \tag{14}$$

Rather than determine the saving of a generation t consumer as the solution to an infinite-horizon maximization problem, I will follow Weil's approach and ask the following question: How much would a generation t consumer save if he earns a wage income w_t, receives a bequest b_{t-1} from his parent, receives gifts totaling ng_{t+1} from his n children, earns a rate of return R_{t+1}, and, in addition, if he is arbitrarily required to leave a bequest of b_t to each of his children and to give a gift of g_t to his parent? Although I cannot answer this question explicitly at this level of generality, the saving function will have the following form

$$s_t = s\big(b_{t-1} - g_t + w_t,\; n(g_{t+1} - b_t), R_{t+1}\big). \tag{15}$$

The saving function in (15) depends on first-period income, second-period income, and the rate of return to saving. Under the assumption that c_{1t} and c_{2t+1} are both normal goods, $s(.,.,.)$ is increasing in its first argument and is decreasing in its second argument. Substituting the competitive factor prices (10, 11) into (15), then substituting the resulting expression into (14) and restricting attention to the steady state yields

$$h(k, b-g) \equiv s\big(b - g + w(k),\; n(g-b), R(k)\big) - nk = 0. \tag{16}$$

I follow Peter Diamond (1965) and confine attention to locally stable steady states (i.e., steady states for which $h_k < 0$). To avoid any complications that may arise from multiple locally stable steady states, I follow Weil and assume that there is a unique locally stable steady state. Let $k = k^*(z)$ be the steady-state capital labor ratio when $b - g = z$.

As a point of reference, consider the steady state of the Diamond (1965) economy in which consumers have neither a bequest motive nor a gift motive. Let k^D denote

[14] Actually, Burbidge departed from the Nash assumption in determining an individual consumer's optimal gift and thus arrived at the analogue of (6′) rather than (6). Under this assumption, the boundary of the admissible region of parameter values is $\alpha\beta = 1$ rather than $\alpha\beta = n$. Adjusting Burbidge's analysis to incorporate the Nash assumption would amend his assumption to $\alpha\beta = n$.

the steady-state capital-labor ratio in the Diamond economy. Because $b = g = 0$ in the Diamond economy, it follows that

$$(17) \qquad k^D = k^*(0).$$

Recall that the saving function $s(\cdot, \cdot, \cdot)$ is increasing in its first argument and is decreasing in its second argument. Therefore, it follows from the definition of $h(k, z)$ in (16) that $h_z(k, z) > 0$ and hence $k^*(z)$ is an increasing function of z.[15] Because $k^{*\prime}(z) > 0$ and $R'(k) < 0$, equation (17) implies that

$$(18) \quad b - g \gtreqless 0 \quad \text{as} \quad k \gtreqless k^D \quad \text{as} \quad R \lesseqgtr R^D.$$

I now present simple conditions which are sufficient for each type of transfer motive to be operative. Essentially, in order for a transfer motive to be operative, it must be sufficiently strong. Proposition 1, which provides a sufficient condition for operative bequests, is due to Weil (1987); Proposition 2, which provides a sufficient condition for operative gifts, is new.

PROPOSITION 1: *If $\beta > n/R^D$, then $b > 0$.*

PROOF:

If $\beta > n/R^D$, then (12) implies that $R^D > n/\beta \geq R$. Therefore, (18) implies that $b - g > 0$, which along with the nonnegativity constraint on g, implies that $b > 0$.

PROPOSITION 2: *If $\alpha > R^D$, then $g > 0$.*

PROOF:

If $\alpha > R^D$, then (12) implies that $R^D < \alpha \leq R$. Therefore, (18) implies that $b - g < 0$, which along with the nonnegativity constraint on b, implies that $g > 0$.

If both transfer motives are sufficiently weak, then there will be no transfers in either direction. Precise conditions are given by

PROPOSITION 3: *If $\beta \leq n/R^D$, $\alpha \leq R^D$, and $\alpha\beta < n$, then $b = g = 0$.*

PROOF:

(by contradiction): Suppose that $b > 0$ so that (13a) implies that $R = n/\beta \geq R^D$. Therefore, (18) implies that $b - g \leq 0$ which implies that $g > 0$. However, if $g > 0$, then (13b) implies that $R = \alpha$, which contradicts the statements above that $R = n/\beta$ and $\alpha\beta < n$. Therefore, $b = 0$. A similar line of argument proves that $g = 0$.

Finally, we consider the case in which $\alpha\beta = n$, which corresponds to the case considered by Burbidge.[16] In general, it is possible for there to be both positive gifts and positive bequests in the steady state. Nevertheless, one can determine whether the net flow of intergenerational transfers is from parents to children ($b - g > 0$), from children to parents ($b - g < 0$), or zero.

PROPOSITION 4: *If $\alpha\beta = n$, then $b - g \gtreqless 0$ as $R^D \gtreqless n/\beta = \alpha$.*

PROOF:

Suppose that $R^D > n/\beta$. It follows from (12) that $R^D > R$ which, according to (18), implies that $b - g > 0$. Similarly, $R^D < \alpha$ implies that $R^D < R$, which according to (18) implies that $b - g < 0$. Finally, $R^D = n/\beta = \alpha$ implies that $R^D = R$, which implies that $b - g = 0$.

The results concerning when the transfer motives will be operative are summarized in Figures 1 and 2. The distinction between Figures 1 and 2 is that the utility function $u(\cdot, \cdot)$ and the production function $f(\ )$ are such that the steady state of the Diamond economy is efficient in Figure 1 but is inefficient in Figure 2. If the Diamond economy is efficient, then Figure 1 indicates that either the gift motive or the bequest motive could be operative; if neither motive is sufficiently strong, then neither motive will be operative. If the Diamond economy is

[15] Formally, $h(k^*(z), z) \equiv 0$, which implies that $k^{*\prime}(z) = -h_z/h_k > 0$.

[16] See fn. 14.

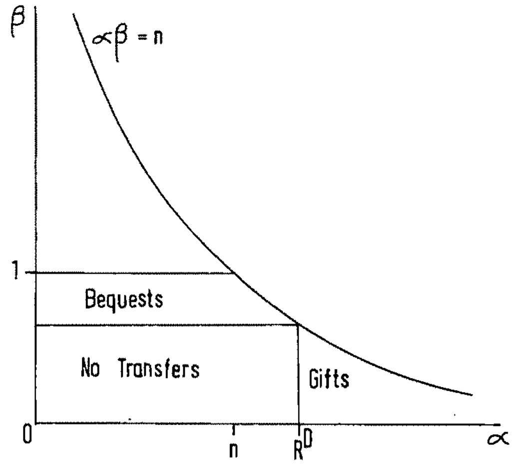


FIGURE 1

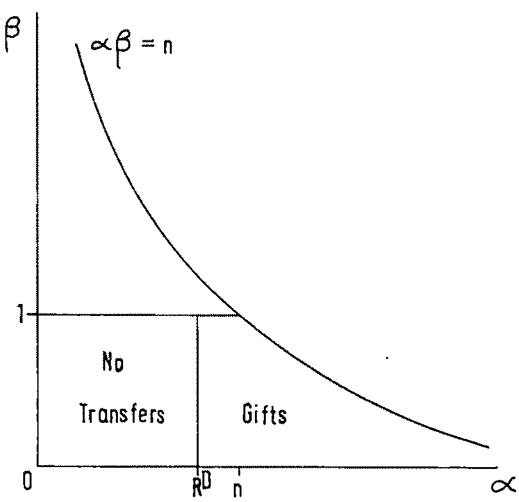


FIGURE 2

inefficient, then Figure 2 indicates that, for admissible values of β, the bequest motive cannot be operative, which is consistent with Weil's (1987) results. However, the gift motive can be operative if it is sufficiently strong. Again, if neither motive is sufficiently strong, then neither will be operative.

The conditions for operative transfer motives are stated in terms of R^D, the steady-state interest rate in the Diamond model. It was Weil's insight to recognize that the R^D provides a useful summary of the utility function $u(\cdot,\cdot)$ and the production function $f(\ )$ for determining whether a transfer motive will be operative. Nevertheless, it would be useful to state the conditions for operative bequests in terms of underlying preferences and technology. As a step toward this goal, I will relate R^D to consumer behavior expressed in terms of the average propensity to consume and to the production function expressed in terms of the capital share of income. Then, for a specific example I will express R^D directly in terms of the parameters of preferences and technology.

Let σ_t denote s_t/w_t, the average propensity to save out of wage income, and let ϕ_t denote the capital share in income, $R_t k_t/y_t$. Because the production function is assumed to be linearly homogeneous, the labor share in income, w_t/y_t, is equal to $1-\phi_t$ so that

$$w_t = [(1-\phi_t)/\phi_t]\, R_t k_t. \tag{19}$$

It follows from (19) and the definition of the average propensity to save, σ_t, that

$$s_t = \sigma_t[(1-\phi_t)/\phi_t]\, R_t k_t. \tag{20}$$

Equating the left-hand side of (14) to the right-hand side of (20) in the steady state of the Diamond economy yields

$$nk^D = \sigma[(1-\phi)/\phi]\, R^D k^D. \tag{21}$$

It follows immediately from (21) that

$$R^D = n\phi/[\sigma(1-\phi)]. \tag{22}$$

It follows from (22) that in the Diamond economy, the steady-state interest rate tends to be large when either the capital share in income, ϕ, is large or the average propensity to save, σ, is small. Of course, the capital share, ϕ, and the average propensity to save, σ, are, in general, endogenously determined. However, there is a special case in which both ϕ and σ are exogenous parameters. If the utility function is logarithmic, $u(c_{1t}, c_{2t+1}) \equiv (1-\sigma)\ln c_{1t} + \sigma \ln c_{2t+1}$, $0<\sigma<1$, then the average propensity to save out of wage income is constant and equal to σ. If

the production function is Cobb-Douglas, $f(k) \equiv Ak^{\phi}$, $0 < \phi < 1$ and $A > 0$, then the capital share in income is constant and equal to ϕ. In this special case, the expression for R^D on the right-hand side of (22) is simply a function of the parameters of preferences and technology. Substituting this expression for R^D in Propositions 1–4 delivers, for this example, a complete characterization, in terms of the parameters of preferences and technology, of situations in which the transfer motives will be operative or inoperative.

V. Concluding Remarks

The effects of changes in the timing of lump-sum taxes depend crucially on whether the motives for intergenerational transfers are operative. In this paper I have derived conditions which determine whether the bequest motive is operative, the gift motive is operative, or neither motive is operative. When neither motive is operative, then changes in the timing of lump-sum taxes affect the intertemporal and intergenerational allocation of resources.

The formal results presented in Propositions 1–4 and summarized in Figures 1 and 2 apply only to the steady state of a representative consumer economy. Future research should be devoted to extending the analysis to the transition path outside the steady state and should analyze economies with interesting heterogeneity. The reason for extending the analysis to the transition path is that the Ricardian Equivalence Theorem requires that all consumers in all generations be linked by operative-intergenerational transfer motives. If some generation has no operative-intergenerational transfer motive, then at least some changes in the timing of lump-sum taxes will affect the intertemporal allocation of resources. The magnitude of the effect would depend on, among other things, the extent and sort of heterogeneity among consumers. For example, heterogeneity with respect to initial wealth or labor income may lead to a situation in which some consumers have operative bequest motives while other consumers in their cohort face binding constraints. In this situation, the Ricardian Equivalence Theorem would not hold; the extent of the departure from the Ricardian Equivalence Theorem, that is, the magnitude of the effect of fiscal policy, would depend on the proportion of consumers who face binding constraints. In a subsequent paper (Abel, 1986), I have begun to explore some of these issues. However, the model in that paper is restricted to Cobb-Douglas technology, logarithmic utility with a bequest motive but no gift motive, and the heterogeneity is restricted to initial wealth. In addition to analyzing more general utility and production functions, future research should analyze the effects of fiscal policy in the presence of heterogeneous labor productivity, secular productivity growth, and two-sided transfer motives.

An additional avenue for future research is to analyze bequest and gift behavior under more general forms of intergenerational transfer motives. Bernheim (1987) has argued that there is no reason to insist on dynamic consistency in modeling the consumption and transfer behavior of families. Recently, Debraj Ray (1987) has examined specifications of intergenerational altruism in which a consumer obtains utility from the utility of many subsequent generations in his family, in addition to obtaining utility directly from his own consumption. If, for example, a consumer cares about his grandchildren's utility in addition to his children's utility and his own consumption, then, in general, the consumption decisions of different generations within the family will display dynamic inconsistency. In addition, Ray has shown that under this sort of altruistic utility function, it is possible for the steady state to be characterized by positive bequests and a dynamically inefficient overaccumulation of capital. The determination of conditions for the bequest motive to be operative or inoperative remains an open question in this more general framework.

REFERENCES

Abel, Andrew B., "Precautionary Saving and Accidental Bequests," *American Economic Review*, September 1985, *75*, 777–91.

______, "An Analysis of Fiscal Policy under

Operative and Inoperative Bequest Motives," mimeo., The Wharton School of the University of Pennsylvania, September 1986, in Elhanan Helpman, Asaff Razin, and Efraim Sadka, eds., *The Economic Effects of the Government Budget*, Cambridge: MIT Press, forthcoming.

Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, *82*, 1095–117.

Bernheim, B. Douglas, "Ricardian Equivalence: Evaluation of Theory and Evidence," in NBER *Macroeconomics Annual*, Stanley Fischer, ed., forthcoming, 1987.

______, **and Bagwel, Kyle,** "Is Everything Neutral? The Implications of Intergenerational Altruism in an Overlapping Generations Model with Marriage," mimeo., Stanford University, November 1984.

Buiter, Willem, "Government Finance in an Overlapping-Generations Model with Gifts and Bequests," in George M. von Furstenberg, ed., *Social Security Versus Private Saving*, Cambridge: Ballinger, 1979.

______ **and Carmichael, Jeffrey,** "Government Debt: Comment," *American Economic Review*, September 1984, *74*, 762–65.

Burbidge, John B. "Government Debt in an Overlapping-Generations Model with Bequests and Gifts," *American Economic Review*, March 1983, *73*, 222–27.

______, "Government Debt: Reply," *American Economic Review*, September 1984, *74*, 766–67.

Carmichael, Jeffrey, "Economic Equilibrium and Steady-State Growth with Intergenerationally-Dependent Preferences," ERP memo. No. 245, Princeton University, 1979.

______, On Barro's Theorem of Debt Neutrality: The Irrelevance of Net Wealth," *American Economic Review*, March 1982, *72*, 202–13.

Cukierman, Alex, "Uncertain Lifetimes and the Ricardian Equivalence Proposition, mimeo., Tel-Aviv University, December 1986.

______ **and Meltzer, Allan,** "A Political Theory of Government Debt and Deficits in a Neo-Ricardian Framework, mimeo., Tel-Aviv University, 1986.

Diamond, Peter A., "National Debt in a Neoclassical Growth Model," *American Economic Review*, December 1965, *55*, 1126–50.

Drazen, Allan, "Government Debt, Human Capital, and Bequests in a Lifecycle Model," *Journal of Political Economy*, June 1978, *86*, 505–16.

Feldstein, Martin S., "Perceived Wealth in Bonds and Social Security: A Comment," *Journal of Political Economy*, April 1976, *84*, 331–36.

______, "The Effects of Fiscal Policies When Incomes are Uncertain: A Contradiction of Ricardian Equivalence," NBER Working Paper No. 2062, November 1986.

Fischer, Stanley, "A Life Cycle Model of Life Insurance Purchases," *International Economic Review*, February 1973, *14*, 132–52.

Friedman, Benjamin M. and Warshawsky, Mark, "The Cost of Annuities: Implications for Saving Behavior and Bequests," mimeo., Harvard University, September 1984.

Gale, Douglas, *Money: in Disequilibrium*, New York: James Nisbet and Company/Cambridge University Press, 1983.

Hori, Hajime and Tsukamoto, Jun, "Voluntary Intergenerational Transfers and the Steady-State Interest Rate," Tohoku University, Discussion Paper No. 59, April 1985.

Kimball, Miles, "Making Sense of Two-Sided Altruism," mimeo., Harvard University, December 1986.

Kotlikoff, Laurence, Shoven, John and Spivak, Avia, "The Impact of Annuity Insurance on Savings and Inequality," presented at the *Conference on the Family and the Distribution of Economic Rewards*, September 20–22, 1984.

Ray, Debraj, "Nonpaternalistic Intergenerational Altruism," *Journal of Economic Theory*, February 1987, *41*, 112–32.

Samuelson, Paul A., "A Turnpike Refutation of the Golden Rule in a Welfare-Maximizing Many-Year Plan," in Karl Shell, ed., *Essays on the Theory of Optimal Economic Growth*, Cambridge: MIT Press, 1967,

269–80.

______, "The Two-Part Golden Rule Deduced as the Asymptotic Turnpike of Catenary Motions," *Western Economic Journal*, March 1968, *6*, 85–89.

Strotz, Robert, "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, 1956, *23*, 165–80.

Tomes, Nigel, "The Family, Inheritance, and the Intergenerational Transmission of Inequality," *Journal of Political Economy*, October 1981, *89*, 928–58.

Weil, Philippe, "'Love Thy Children': Reflections on the Barro Debt Neutrality Theorem," *Journal of Monetary Economics*, May 1987, *19*, 3, 377–91.

A Model of Medieval Grain Prices: Comment

By B. TAUB*

A dynamic model of grain storage supports and elaborates the McCloskey/Nash empirical findings about grain price movements and interest rates. In particular, their finding of a negative time derivative of the growth rate of grain prices is explained, and a means of econometrically distinguishing the interest rate from storage rents is stated.

Donald McCloskey and John Nash (1984) estimated medieval interest rates from seasonal grain price changes. They argued that the purchase of grain for storage must be compensated by interest when it is resold, and so in an intertemporally arbitraged market, the price of grain must rise at the rate of interest. Since depreciated grain represents foregone value, the interest rate is gross of depreciation. They found this gross interest rate to be high.

This paper supports their conclusions with a theoretical model. In particular, McCloskey and Nash's decision to examine the price rises only in the post-harvest period is shown to be sound, and their finding of a negative time derivative of the growth rate of grain prices is explained. Their conclusions about the separability of costs of storage from interest rate effects, however, are shown to be erroneous, and so their interest-rate estimates are not econometrically identified, and must be considered upper bounds at best.

A dynamic model of seasonal crop production and storage is presented in Section I. Section II discusses the dynamics and proposes a solution of the identification problem.

I. Tastes and Technology

McCloskey and Nash assumed that the planting and harvest occurred at fixed times in the season, and that transportation costs were so high that there was no arbitrage of grain between regions with different harvest dates. Such an economy is properly modeled with the price of grain dependent only on local conditions, that is, by a closed-economy model.

Individuals are assumed identical.[1] They plant in the spring, consume from their stored grain during the summer, then harvest at summer's end. The harvest goes into storehouses and provides consumption until the following harvest. Individuals thus solve three problems: how much to plant, how much to consume from inventory between planting and harvest, and how much to consume from inventory between harvest and planting. They therefore solve the following problem at planting time:

$$(1)\quad V(G_0) = \max_{P_0, \{c(\tau)\}} \left\{ \int_0^{T_P} e^{-\rho\tau} U(c(\tau))\, d\tau + e^{-\rho T_P} V(s(T_P)) \right\}$$

*Assistant Professor, Department of Economics, Virginia Tech., Blacksburg, VA 24061. I thank John Nash for correspondence and discussion that stimulated this paper and a referee for several helpful suggestions. Ronald Michener provided a useful comment on an earlier version of the paper.

[1] The assumption that individuals are identical greatly simplifies the equilibrium analysis. This is a standard technical tactic in dynamic general equilibrium models where no externalities are present (as here), because it allows one to read off prices from intertemporal marginal rates of substitution that one has found by solving a maximum problem. Robert Lucas (1978) is but one example where this is done.

subject to the following constraints:

$$(2)\qquad \dot{s} = -(c + \delta s + \phi(s))$$

$$(3)\qquad \lim_{t \downarrow T_1} s(t) = s(T_H) + f(P_0)$$

$$(4)\qquad s(t) \geq 0$$

$$(5)\qquad s(0) = G_0 - P_{0'}$$

$$(6)\qquad G_0 \text{ given,}$$

where $c(t)$ = grain consumption at time t; $s(t)$ = stock of grain in storage at time t; ρ = discount factor; δ = depreciation rate of stored grain, assumed constant; T_H = time of harvest; T_P = next planting time, $T_H < T_P$; G_0 = initial stock of grain; P_0 = initial quantity of grain planted; $U(\cdot)$ = utility function; $V(\cdot)$ = value of the stock of grain available at the next planting time; $f(\cdot)$ = production function; and $\phi(\cdot)$ = cost function.

The utility, production, and cost functions have the usual properties: they are differentiable, the utility and production functions are increasing and concave, and the cost function is increasing and convex. The cost function can be thought of as the depreciation resulting from storage; its key aspect is that it is an increasing function of the quantity of stored grain. The value function, $V(\cdot)$, will have the properties of a utility function; since the problem is repeated anew each season, $V(\cdot)$ is endogenous and its properties must be derived. This is done in the Appendix.

The problem (1) requires choosing both the continuous-time path of optimal consumption and the annual discrete choice of planting. The solution of (1) is facilitated by solving a pair of continuous-time subproblems with optimal control methods and then linking them in a discrete dynamic programming problem.

First, the continuous-time problem between harvest and replanting can be expressed as

$$(7)\quad W(s(T_H) + H(T_H)) = \max_{\{c(t)\}} \left\{ \int_{T_H}^{T_P} e^{-\rho\tau} U(c(\tau))\, d\tau + e^{-\rho(T_P - T_H)} V(s(T_p)) \right\}$$

subject to

$$(8)\qquad s(T_H) + H(T_H) \text{ given,}$$

and to (2) and (4), where $H(T_H)$ denotes the harvest at time T_H. This is a standard control-theory problem with a terminal value.[2] The current-value Hamiltonian is

$$\mathscr{H}(c, s) = U(c(\tau)) - q(c + \delta s + \phi(s)).$$

The necessary conditions for an optimum are

$$(9)\qquad \dot{q} = (\rho + \delta + \phi'(s))q$$

$$(10)\qquad q(t) = U'(c(t))$$

$$(11)\qquad q(T_P) = V'(s(T_P)).$$

From (10) it can be seen that $q(t)$ has the usual interpretation as the shadow price of the state variable, so in this case we will interpret it as the price of grain. In a competitive equilibrium model, $q(t)$ would be the market price of grain. For the optimal path of consumption $c^*(t)$, we can define the intermediate value function

$$(12)\quad J_{T_P - T_H}(s(T_H) + H(T_H)) = \int_{T_H}^{T_P} e^{-\rho(\tau - T_H)} U(c^*(\tau))\, d\tau,$$

and therefore

$$(13)\quad W(s(T_H) + H(T_H)) = J_{T_P - T_H}(s(T_H) + H(T_H)) + e^{-\rho(T_P - T_H)} V(s(T_P)).$$

The next subproblem is to choose the path of consumption between planting and

[2]See Morton Kamien and Nancy Schwartz (1981), pp. 143–50.

harvest:

$$(14)\quad \max_{\{c(\tau)\}}\left\{\int_0^{T_H} e^{-\rho\tau}U(c(\tau))\,d\tau + e^{\rho T_H}W(s(T_H)+H(T_H))\right\},$$

subject to (2), (4), and with $s(0)$ and $H(T_H)$ given. Proceeding as in the previous problem, the necessary conditions are

$$(15)\qquad \dot{q} = (\rho+\delta+\phi'(s))q,$$

$$(16)\qquad q(t) = U'(c(t)),$$

$$(17)\qquad q(T_H) = W'(s(T_H)+H(T_H)).$$

Note that the last condition provides the missing boundary condition for the "winter" problem, (7).

We can now piece together the winter and summer problems to state the problem of optimal planting. As in (12), define the intermediate value

$$J_{T_H}(s(0)) = \int_0^{T_H} e^{-\rho\tau}U(c^*(\tau))\,d\tau.$$

Noting that $H(T_H) = f(G_0 - s(0))$, we have

$$(18)\quad V(G_0) = \max_{s(0)}\left\{J_{T_H}(s(0)) + e^{-\rho T_H}W(s(T_H)+f(G_0-s(0)))\right\}.$$

As in typical dynamic programming problems, the marginal value of the state variable equals the marginal utility of consumption. Therefore, conditions (10), (16), and (17), together with $U'(c(T_H)) = W'(s(T_H) + H(T_H))$ imply that q does not jump at planting time.

The final stage of the analysis is to prove the existence and properties of the value function, $V(\cdot)$, with dynamic programming methods. This is secondary to the analysis of the seasonal dynamics, so it is relegated to the Appendix.

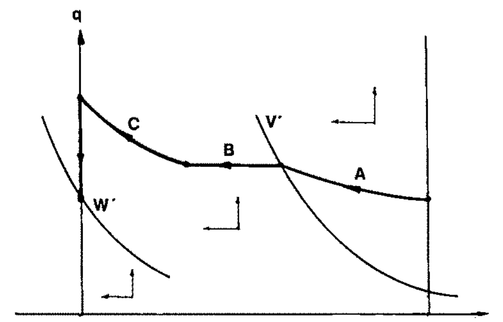


FIGURE 1. PHASE DIAGRAM FOR STOCK OF STORED GRAIN (s) AND SHADOW PRICE OF GRAIN (q)

II. Dynamic Behavior

A phase diagram can now be used to analyze the dynamics. From (2) and (9), it is apparent that grain stocks continually shrink, while their shadow price continually rises. Figure 1 shows that during the winter segment (segment A), stored stocks are being drawn down, while consumption, which is inversely related to the shadow price of grain, continually falls. At planting time, the stock of stored grain falls by a discrete amount, but the shadow price does not shift (segment B). Consumption and the stored stock continue to fall until harvest (segment C).

An arbitrage argument shows that all of the grain will be drawn out of storage just before the harvest under perfect foresight. If some quantity remains, then speculators would "go short"; that is, they would borrow grain before harvest, sell it for use in consumption, then repay the loan after harvest when its consumption value is low. This tends to reduce the stock before harvest and raise its price. The impossibility of transporting the grain backward in time from the harvest if pre-harvest stocks are not carried over prevents the post-harvest price from rising in response to this arbitrage.

Since the marginal value of stored grain is not infinite just before harvest, that is, $W'((s(T_H)+H(T_H)) < \infty$ for $s(T_H) = 0$, the terminal point of the optimal summer path (segment C of Figure 1) occurs at zero-stored grain but with a finite shadow price.

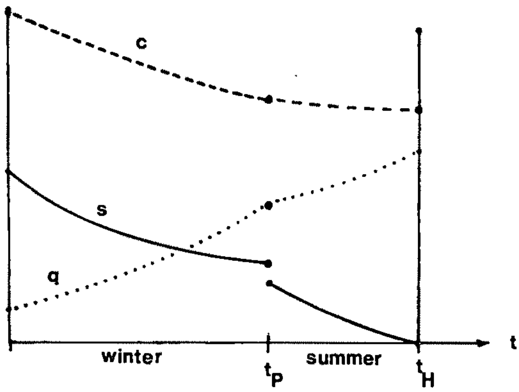


FIGURE 2. PATHS OF CONSUMPTION (c), STOCK OF STORED GRAIN (s), AND SHADOW PRICE OF GRAIN (q) OVER THE SEASON

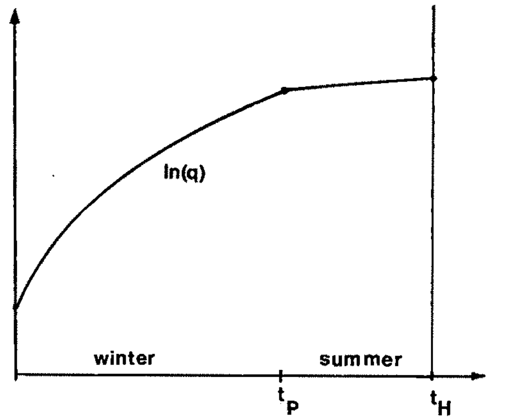


FIGURE 3. PATH OF THE LOGARITHM OF THE SHADOW PRICE OF GRAIN OVER THE SEASON

It is now possible to characterize the path of the grain stock, consumption, and price in more detail. Figure 2 shows their seasonal path. The shadow price drops after harvest, then climbs at a rate faster than the interest rate, $\rho + \delta$, because of the presence of instantaneous storage costs. Thus any attempt to impute the gross interest rate from the price path would find only an upper bound.[3] The instantaneous storage cost was and is likely to be small, representing variable costs of the storage of the gross stock of grain. They are, however, likely to be an increasing function of that stock. In a market economy in which commercial barns and silos were used for storage, the efficient ones would be filled first; efficiency here means those in which grain depreciation is lowest. With these filled, farmers would have resorted to high-depreciation storage, such as smaller bins and even open ground. The efficient barns would then receive a rent equal to the difference of their marginal cost in depreciation and that of the least efficient storehouse in use.

Inspection of (9) and (15) shows that as grain is withdrawn from high-cost storehouses over the season, the percentage rate of growth of its price should decline. The declining marginal cost of storage as grain is withdrawn means that the path of the logarithm of the shadow price of grain will have a negative second derivative (see Figure 3). This is precisely the pattern found by McCloskey and Nash (see their Table 2 and associated discussion). There will be a kink at the time of planting due to the discrete fall in marginal storage costs.

A. *Uncertainty*

Under perfect foresight, arbitrage prevents any price decline except at harvest time. If the outcome of planting is uncertain, price declines can occur before harvest as McCloskey and Nash speculate. Suppose there are just two possible outcomes of the harvest, good and bad, and that the outcome is revealed midway between planting and harvest. Excess grain will be stored at planting time as insurance against a bad harvest. When news of good harvest arrives, the insurance stock can be released for consumption. The result is that the optimal path will jump down in midsummer, and the rate of grain consumption will rise in order to use up the insurance stock before harvest. If the uncertainty about the harvest is dispelled gradually over the whole growing season, then the grain price will fall continuously.

[3] The model here does not capture the effects of a long-run appreciation of grain prices, which would also cause the overestimation of the interest rate.

It is the fear of a bad harvest that induces the holding of the insurance stocks. If bad harvests occur only once in ten years on average, then the pre-harvest price fall will be the norm; only in bad years will the price continue to rise through the growing season. Since the pattern of prices in the growing season reflects stochastic events more than saving behavior, McCloskey and Nash properly ignored these periods in estimating the interest rate.

B. *Multiple Grains and Identification*

If there are two grains, say wheat and rye, then their price behavior can be combined to reveal the true interest rate. Suppose wheat and rye have the same planting and harvest times,[4] and are similar enough to be stored in the same warehouses, but rye is inferior in consumption. The cost of storage would then be a function of the sums of the stocks of grain, since a bushel of rye "crowds out" a bushel of wheat. Choosing wheat as the numeraire in which storage costs are computed, the objective is then

$$\text{(19)} \qquad \max \int_0^{T_P} e^{-\rho\tau} U(c_W, c_R)\, d\tau,$$

subject to

$$\text{(20)} \quad \dot{s}_W = -\left(c_W + \delta s_W + \phi(s_W + s_R)\right),$$

$$\text{(21)} \quad \dot{s}_R = -\left(c_R + \delta s_R\right).$$

The costate equations are then

$$\text{(22)} \quad \dot{q}_W = \left(\rho + \delta + \phi'(s_W + s_R)\right) q_W,$$

$$\text{(23)} \quad \dot{q}_R = (\rho + \delta) q_R + \phi'(s_W + s_R) q_W.$$

Subtracting these equations eliminates the storage cost, leaving

$$\text{(24)} \quad (\dot{q}_W - \dot{q}_R) = (\rho + \delta)(q_W - q_R).$$

The percentage growth in the difference of the prices then reveals the underlying discount rate. Note that the percentage growth of the difference is not equal to the difference of the percentage growth rates. This constrains the econometric implementation to price pairs of the two grains on matching series of dates.

[4] If planting and harvest times were different, the extremes of the harvest cycle would be smoothed and the storehouses would be used more efficiently. The basic condition derived here, (24), would remain intact between adjacent planting and harvest dates, however.

Appendix

In this appendix the properties of continuous-time optimal control and discrete dynamic programming are combined to show that the value function, $V(\cdot)$, exists and is increasing and concave.

PROPOSITION 1: *Define*

$$K(s(0)) = J_{T_H}(s(0)) + J_{T_P - T_H}\left(s(T_H) + H(T_H)\right).$$

Then K is an increasing function.

PROOF:

Refer to the phase diagram in Figure 1. Consider two initial stocks, $s^0(0)$ and $s^1(0)$, with $s^0(0) < s^1(0)$. The winter path for $s^1(0)$ (segment A^1), of duration $T_P - T_H$, must lie below the corresponding path for $s^0(0)$, for otherwise the V' terminus is not attained within time $T_P - T_H$.

Suppose that the same amount of planting is done for both $s^0(0)$ and $s^1(0)$. Then the segment B^0 will have the same length as B^1, and segments C^0 and C^1 will both terminate on the same function $W'(s(T_H) + H(T_H))$. But this requires that segment C^1 terminate with positive stocks remaining. Increased planting lowers the W' function so that segment C^1 terminates at zero. This leaves the C^1 segment below the C^0 segment. Thus the optimal s^1 path is therefore everywhere below the s^0 path, corresponding to higher consumption everywhere on the path. Thus $K(s^1(0)) > K(s^0(0))$. This completes the proof.

PROPOSITION 2: *$K(s)$ is concave.*

PROOF:

Choose α, $0 < \alpha < 1$, and define $s^\alpha \equiv \alpha s^0 + (1-\alpha)s^1$. The associated optimal consumption paths for s^0 and s^1 are c^0 and c^1.

Then

$$\alpha K(s^0)+(1-\alpha)K(s^1)$$
$$=\int_0^{T_P} e^{-\rho\tau}\{\alpha U(c^0(\tau))+(1-\alpha)U(c^1(\tau))\}\,d\tau$$
$$\leq \int_0^{T_P} e^{-\rho\tau}U(\alpha c^0(\tau)+(1-\alpha)c^1(\tau))\,d\tau,$$

with the equality following from the definition of K and the inequality following from the concavity of $U(\cdot)$. Define the path $c^\alpha \equiv \alpha c^0+(1-\alpha)c^1$. Then c^α is a feasible path for s^α since it exactly exhausts the initial stock. It is not necessarily true that c^α is the optimal path for s^α, and so

$$\int_0^{T_P} e^{-\rho\tau}U(c^\alpha(\tau))\,d\tau \leq K(s^\alpha).$$

This completes the proof.

PROPOSITION 3: *A unique value function exists, which is concave and increasing.*

PROOF:

V is defined by

$$\text{(A1)}\quad V=\max_{s(0)}\{K(s(0))+e^{-\rho T_P}V(s(T_P|s(0)))\}$$

subject to

$$\text{(A2)}\qquad s(0)\leq G_0.$$

We first demonstrate that the operator TV defined by (A1) is a contraction. The following properties hold for T:

(*i*) *Monotonicity*. Suppose $V^2(x)>V^1(x)$ for all x. Then

$$TV^1(G)=\max_{s(0)}\{K(s(0))+e^{-\rho T_P}V^1(s(T_P))\}.$$

Define the optimal consumption path by $c^1(s(0))$, and the resulting terminal stock by $s^1(T_P)$. We have the following inequalities:

$$TV^1=\int_0^{T_P'} e^{-\rho\tau}U(c^1(\tau))\,d\tau+e^{-\rho T_P}V^1(s^1(T_P))$$
$$\leq \int_0^{T_P} e^{-\rho\tau}U(c^1(\tau))\,d\tau+e^{-\rho T_P}V^2(s^1(T_P))$$
$$\leq \int_0^{T_P} e^{-\rho\tau}U(c^2(\tau))\,d\tau+e^{-\rho T_P}V^2(s^2(T_P))=TV^2.$$

The first inequality follows from the hypothesis and the second from the optimality of the c^2 path for V^2.

(*ii*) *Discounting*. If A is a constant,

$$T(V+A)$$
$$=K(s(0))+e^{-\rho T_P}(V(s(T_P)+A)$$
$$\leq K(s(0))+e^{-\rho T_P}V(s(T_P))+e^{-\rho T_P}A$$
$$=TV+e^{-\rho T_P}A.$$

The final equality follows because the terminal condition is not affected by the constant; that is, $q(T_P)=d/ds(V(s(T_P))+A)=V'(s(T_P))$.

By David Blackwell's theorem (1965), these two properties are sufficient for T to be a contraction mapping, and so a unique V exists. The proof that V is increasing and concave now follows from the properties of $K(\cdot)$ demonstrated in the preceding propositions in combination with standard dynamic programming methods; see, for example, R. E. Lucas (1978, p. 1432). This completes the proof.

REFERENCES

Blackwell, David, "Discounted Dynamic Programming," *Annals of Mathematical Statistics*, 1965, *36*, 226–35.

Kamien, Morton and Schwartz, Nancy, *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*, New York: North-Holland, 1981.

Lucas, Robert E. Jr., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, 1429–45.

McCloskey, Donald and Nash, John, "Corn at Interest: The Extent and Cost of Grain Storage in Medieval England," *American Economic Review*, March 1984, *74*, 174–87.

The Dynamics of Population Growth, Differential Fertility, and Inequality: Note

By C. Y. CYRUS CHU*

In a recent article in this *Review* (1986), David Lam made an interesting attempt to examine the relationship between population growth and the distribution of income. In Sections I and II of his paper, in which no income mobility was allowed, Lam successfully analyzed the effects of adding to the economy a group of immigrants (with income distribution different from the original residents) on two inequality measures. In Section III, in which mobility across income groups was taken into consideration, Lam claimed that "*if $M_{1i} < M_{11}$ $\forall i, \ldots,$ then an increase in the fertility of the poor will unambiguously increase the percent poor in the steady state*" (p. 1110), where M_{ij} is the probability that a child of class j becomes a member of class i. This is a rather strong proposition that says, in effect, in order to tell whether an increase in the ith-income groups's fertility would increase or reduce the steady-state proportion of the ith-income group, all the information we need is the ith row of the mobility matrix.

The purpose of this note is to call into question Lam's proposition and to demonstrate that the condition $M_{11} > M_{1i}$ $\forall i$ alone is not sufficient to predict the rise or fall of the percent poor in the steady state as a result of an increase in the fertility of the poor. I will first give a counterexample in Section I. In Section II, I provide a correction for the error that is made in the derivation of Lam's proposition.

I. A Counterexample

Following Lam's notations, denote the size of n-income classes in period t by $P_t' = [\mathbf{P}_{1,t}, \mathbf{P}_{2,t}, \ldots, \mathbf{P}_{n,t}]$, and the income-specific net reproduction rates by a diagonal matrix $\mathbf{F}$. Let $\mathbf{M}$ be an intergenerational mobility matrix, where element $\mathbf{M}_{ij}$ is the probability that a child of class j becomes a member of class i. Thinking of each period as a generation, Lam showed that the distributions of income between two generations would be characterized by the following identity (equation (8), p. 1109):

$$\mathbf{P}_t = \mathbf{MFP}_{t-1}. \tag{1}$$

Let the proportion of the population in income class i at time t be $\pi_{i,t}$. Dividing both sides of (1) by $N_{t-1} \equiv \Sigma_{i=1}^{n} P_{i,t-1}$, the total population size at time $t-1$, yields

$$\begin{aligned}\mathbf{MF}\pi_{t-1} &= \mathbf{MF}(\mathbf{P}_{t-1}/N_{t-1}) \\ &= \mathbf{P}_t/N_{t-1} = (\mathbf{P}_t/N_t)g_t = \pi_t g_t,\end{aligned} \tag{2}$$

where $g_t \equiv (N_t/N_{t-1})$ is the population growth rate at period t, and $\pi_t' \equiv [\pi_{1,t}, \pi_{2,t}, \ldots, \pi_{n,t}]$. In the steady state, we drop the time subscripts in (2) and rewrite it as

$$\mathbf{MF}\pi = \pi g. \tag{3}$$

For demonstration purposes, consider the case that $n = 3$. Then (3) can be explicitly written as

$$\mathbf{F}_1\mathbf{M}_{11}\pi_1 + \mathbf{F}_2\mathbf{M}_{12}\pi_2 + \mathbf{F}_3\mathbf{M}_{13}\pi_3 - g\pi_1 = 0 \tag{4}$$

$$\mathbf{F}_1\mathbf{M}_{21}\pi_1 + \mathbf{F}_2\mathbf{M}_{22}\pi_2 + \mathbf{F}_3\mathbf{M}_{23}\pi_3 - g\pi_2 = 0 \tag{5}$$

$$\mathbf{F}_1\mathbf{M}_{31}\pi_1 + \mathbf{F}_2\mathbf{M}_{32}\pi_2 + \mathbf{F}_3\mathbf{M}_{33}\pi_3 - g\pi_3 = 0, \tag{6}$$

*Graduate School of Business, National Taiwan University and Institute of Economics, Academia Sinica, Nankang, Taipei, Taiwan. I am indebted to S. K. Hsu and two anonymous referees for helpful suggestions.

where $\mathbf{F}_i$ is the ith diagonal term of $\mathbf{F}$. In (4)–(6), since $\pi_1 + \pi_2 + \pi_3 = 1$, the degree of freedom of the vector π is two. Thus, g plus any two elements of $\{\pi_1, \pi_2, \pi_3\}$ are the three endogenous variables to be determined by the system (4)–(6). Consider the case

$$\mathbf{M} = \begin{bmatrix} .3 & .29 & 0 \\ 0 & .6 & .5 \\ .7 & .11 & .5 \end{bmatrix}.$$

Clearly, **MF** is positively regular and the condition $\mathbf{M}_{11} > \mathbf{M}_{1i}$, $i = 1,2,3$ is satisfied. Let the values of $(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3) = \tilde{\mathbf{F}}$ change from $\tilde{\mathbf{F}}_0 = [1,1,1]$ to $\tilde{\mathbf{F}}_1 = [1.01, 1, 1]$. Then the values π_1, π_2, π_3, and g can be solved from (4)–(6). The result turns out to be: When $\tilde{\mathbf{F}} = \tilde{\mathbf{F}}_0$, $(\pi_1, \pi_2, g) = (.1870968, .4516129, 1)$, and when $\tilde{\mathbf{F}} = \mathbf{F}_1, (\pi_1, \pi_2, g) = (.1870272, .4507144, 1.00187)$. Since π_1 decreases as $\mathbf{F}_1$ increases, this result is clearly in contradiction with Lam's claim.

II. Correct Comparative Statics

Since Lam's conclusion was derived from equations (9) and (10) of his paper (p. 1109), it is only logical that we take a detailed look at Lam's derivation after it has been contradicted by my counterexample. When **M** is 3∗3, we can divide both sides of his equation (9) (with the help of the identities $N_t/N_{t-1} = g_t$ and $P_{i,t-1}/N_{t-1} = \pi_{i,t-1}$), to derive

$$(7) \quad \pi_{1,t} g_t = \pi_{1,t-1}\mathbf{F}_1\mathbf{M}_{11} + \pi_{2,t-1}\mathbf{F}_2\mathbf{M}_{12} + \pi_{3,t-1}\mathbf{F}_3\mathbf{M}_{13}.$$

Since at time $t-1$, $\pi_{i,t-1}$ is known for all i, we can take differentiation on both sides of (7) and use the relationship $dg_t = d(\Sigma\pi_{i,t-1}\mathbf{F}_i)$ to derive Lam's equation (10) (which appears to be missing a positive constant). Thus, Lam's result is correct for a *single period analysis*. However, it is noticed from equation (2) that a change in $\mathbf{F}_1$ at period $t-1$ will also change $\pi_{2,t}$ and $\pi_{3,t}$, which, in turn, will affect $\pi_{1,t+1}$, $\pi_{2,t+1}$, and g_{t+1}. But this cross-equation *indirect effect* would never have a chance to appear in a single period analysis, and hence any conclusion about the steady state (which essentially endogenizes all the indirect effect in subsequent periods) drawn from this single period analysis may be misleading. In order to find out the effect of an increase in $\mathbf{F}_1$ on the steady-state π_1, one should take differentiation on the system (4)–(6). For demonstration purposes, let us set

$$\mathbf{M} = \begin{bmatrix} a & c & 0 \\ 0 & e & 1-b \\ 1-a & 1-c-e & b \end{bmatrix}.$$

This matrix clearly satisfies the conditions of positive regularity as long as all of the nonzero elements are positive. Furthermore, if $a > c$, then the condition $\mathbf{M}_{11} > \mathbf{M}_{1i}$, $i = 1,2,3$, is also satisfied. With this specification, we can totally differentiate equations (4)–(6) and get

$$[d\pi_1, d\pi_2, dg]A' = [-a\pi_1 d\mathbf{F}_1, 0, -(1-a)\pi_1 d\mathbf{F}_1],$$

where

$$A = \begin{bmatrix} \mathbf{F}_1 a - g & \mathbf{F}_2 c & -\pi_1 \\ -\mathbf{F}_3(1-b) & \mathbf{F}_2 e - g - \mathbf{F}_3(1-b) & -\pi_2 \\ \mathbf{F}_1(1-a) - \mathbf{F}_3 b + g & \mathbf{F}_2(1-c-e) - \mathbf{F}_3 b + g & -(1-\pi_1-\pi_2) \end{bmatrix}.$$

Thus, the comparative static result is

$$(8) \quad \frac{\partial \pi_1}{\partial \mathbf{F}_1} = \frac{-\pi_1}{|A|}\left[-\pi_2(a\mathbf{F}_3 - \mathbf{F}_2(a-c)) - (a-\pi_1)(\mathbf{F}_2 e - g - \mathbf{F}_3(1-b))\right].$$

In the example given in Section I, we set $\{a, b, c, e\} = \{.3, .5, .29, .6\}$ (note that $a > c$ in this case), and the original situation is assumed to be $\mathbf{F}_1 = \mathbf{F}_2 = \mathbf{F}_3 = 1$. Then clearly $g = 1$ and $|A|$ reduces to

$$|A| = [(1-.5) + (1-.6)] \quad (.3-1) - .29(1-.5) < 0,$$

which in turn implies that

$$\operatorname{sgn}[\partial_{\pi 1}/\partial \mathbf{F}_1] = \operatorname{sgn}[-\pi_2(a\mathbf{F}_3 - \mathbf{F}_2(a-c)) - (a-\pi_1)(\mathbf{F}_2 e - g - \mathbf{F}_3(1-b))].$$

Furthermore, it has been shown in Section I that the solution of (π_1, π_2, g), given the above specification, is $(\pi_{1*}, \pi_{2*}, g_*) = (.1870968, .4516129, 1)$. With this information, we can see that the sign of the comparative statics around the neighborhood of $(\pi_{1*}, \pi_{2*}, g_*)$ will be

$$\begin{aligned}\operatorname{sgn}[\partial \pi_1/\partial \mathbf{F}_1] &= \operatorname{sgn}[-\pi_{2*}(.3-(.3-.29)) - (.3-\pi_{1*})(.6-1-.5)]\\ &= \operatorname{sgn}[-.0293548] < 0.\end{aligned}$$

The upshot: Knowing $\mathbf{M}_{1i} < \mathbf{M}_{11} \forall i$ is not enough to tell whether an increase in the fertility of the poor will increase the percent poor in the steady state; rather it is the values of all terms in **M** and **F** that hold the key. Although Lam did ground-breaking work in highlighting the importance of the relationship between population growth and income distribution, the above exposition clearly demonstrates that this relationship is even more complex than he recognized.

REFERENCE

Lam, David, "The Dynamics of Population Growth, Differential Fertility, and Inequality," *American Economic Review*, December 1986, *76*, 1103–16.

The Competitive Effects of Vertical Agreements: Comment

By G. FRANK MATHEWSON AND RALPH A. WINTER*

Recent economic analyses of vertical restraints and integration emphasize the circumstances under which these arrangements are socially efficient. Efficiency claims have proven contentious, however, for exclusive dealing, a vertical restraint that prohibits any outlet carrying a manufacturer's product from stocking substitute brands. This paper analyzes the impact of exclusive dealing on competition and allocative efficiency.

Opinions expressed on the impact of exclusive dealing range from the extreme view that it is invariably anticompetitive, to the view that it is always procompetitive. The U.S. Supreme Court has concluded that in "hundreds if not thousands of communities" where there is a single retailer for a product, exclusive dealing leads to a monopoly position for the single-dominant manufacturer, thus eliminating competition (*Standard Fashion v. Magrane-Houston Co.*, 1922). Robert H. Bork (1978) disagrees. Taking the extreme case of one retail outlet market, Bork argues that the retailer acts as an agent for the consumer. A manufacturer would have to bribe the single retailer to carry its product exclusively and could do so only by lowering its wholesale price. The retailer would accept the exclusive arrangement only if the reduction in price more than compensated for the reduction in the consumer's choice set. Bork concludes that allowing exclusive dealing can only increase competition and benefit consumers.

William S. Comanor and H. E. Frech (CF, 1985) attempt to determine the competitive impact of exclusive dealing in a formal model. CF argue that exclusive dealing is anticompetitive, and should be disallowed, when it raises the costs of entry into a market for a rival firm.

The impact on rivals' entry costs is the wrong criterion for assessing the efficiency of a trade practice. The criterion that economists accept for evaluating antitrust policy is the maximization of efficiency or welfare. In arguing for a rule of reason approach to exclusive dealing, CF state explicitly that they "do not consider ultimate welfare gains or losses" (CF, p. 539).

We examine exclusive dealing in a model in which two manufacturers sell to a large number of retailers, each of which has a local monopoly in selling to final consumers. Two critical aspects of retail sectors drive our results. A typical retailer has a small share of the wholesale market as a buyer; but retailers often have some local price-setting power as sellers because of uniqueness of locations and local brand names.

This simple model leads to a tradeoff in the impact of exclusive dealing on prices and welfare. Without exclusive dealing, a representative retailer sells the products of all (both) manufacturers. Exclusive dealing imposed by the dominant manufacturer eliminates its rival from the market—exclusive dealing both reduces actual competition and restricts the consumers' choice set. But under exclusive dealing, manufacturers compete on the basis of wholesale prices for the right to be selected by the retailer. *Potential competition replaces actual competition as the disciplining force in the market.*

The tradeoff between actual competition without exclusive dealing and potential competition with exclusive dealing means that the impact on retail price and welfare of allowing exclusive dealing is ambiguous in our model. The practice need not be efficient as Bork argues. In contrast to CF, however, exclusive dealing may increase welfare even when it leads to the complete monopolization

*Departments of Economics, University of Toronto, Toronto, Canada M5S 1A1, and the Hoover Institution and the University of Toronto, Toronto, Canada, M5S 1A1, respectively. We wish to thank Yehuda Kotowitz, George Stigler, and two referees for helpful comments, and the Social Sciences and Humanities Research Council of Canada and the Center for the Study of the Economy and the State at the University of Chicago for funding.

of a retail sub-market. The key is that the manufacturer imposing exclusive dealing may capture the market only by lowering the wholesale price, which indirectly lowers the retail price. Potential competition under exclusive dealing may be stronger in disciplining the dominant manufacturer than actual competition when exclusive dealing is prohibited—so much so that a drop in price with exclusive dealing more than offsets the welfare loss from the reduction in the set of products available to consumers.

I. The Model

We analyze the private and social incentives for exclusive dealing in the simplest possible model. Two producers sell substitute products to a large number of retailers, each of whom has a local monopoly over a subset of consumers.

The contract in the wholesale market is determined by the following game. The producers offer wholesale contracts to the retailers—each contract specifies a wholesale price and possibly an exclusive dealing requirement. We assume that producers' decisions on exclusive dealing (ED) are taken simultaneously and prior to their decisions on wholesale price offers. Each retailer chooses the contract that yields the higher retail profits (if one of the contracts offered contains an ED requirement) or purchases both products at the specified wholesale prices if no ED requirement is specified; the retailer then sells to the retail consumers in its local market. The equilibrium concept employed is the usual perfect Nash equilibrium.

Five assumptions embedded in this simple structure require elaboration. First, an explanation of ED requires some price-setting power in the retail sector. Following Bork (1978, p. 307), we take the extreme case of each retailer selling in a one-store town—a retail sub-market where the fixed costs are sufficiently large that only one outlet is supported.[1] This corresponds, for example, to the *Standard Fashion* case of 1922, in which much of the retail sales came from one-store rural markets. That ED may increase competition in the market, even when it leads to complete *ex post* exclusion of the rival firm, is most clearly illustrated under the assumption of single-store markets but would only be strengthened when a rival can retain some market share.

The second assumption implicit in the specification is that the retailer, a complete monopolist, has no *monopsony* power.[2] This is the standard economic assumption that in a market with a large number of buyers and one or two sellers, the market prices (contracts) are determined by the sellers. This assumption contrasts with the analysis of Bork who takes the position that "[the retailer], not Standard [the manufacturer], is in the better position to bargain. The retailer has alternative suppliers. Standard has no alternative outlet with which to reach customers in that town" (Bork, p. 307). Bork's assumption is critical in his procompetitive assessment of exclusive dealing.

The idea that local monopoly power in the retail market translates to monopsony power in the wholesale market is intuitive, popular, and wrong. That retailers typically have a small share of the wholesale market suffices to justify the assumption that they take contracts as given, irrespective of their exclusive reselling rights. In addition, the assumption of many retailers means that the costs of coordinating buying through a retailer cartel, to extract monopsony power, would be prohibitive.

[1] The retail monopoly assumption corresponds to the CF restriction of a prohibitive cost disadvantage to independent retailing by the rival (a very high entry barrier). The advantage of our assumption is that it permits a fully specified model.

[2] This means that the retailer takes the best contract offered in the wholesale market rather than setting the contract just acceptable to the manufacturer. This is justified by the interpretation that the retailer is one of "hundreds if not thousands" of retailers in segmented markets (towns), each of a very small size compared with the entire wholesale market for either manufacturer's product. The theoretical explanation of why it is empirically reasonable to impute zero-monopsony power to buyers with small market shares is an open issue not explored here.

A third assumption in our model, captured with simultaneous contract offers, is the contestability condition that neither manufacturer incurs any sunken costs in offering a contract. In the *Standard Fashion* case, for example, the contracts ran for two years with half of the contracts up for renewal each year. This represents substantial competition for the favors of retailers.

A fourth assumption is that the decisions on ED are made conditional on both decisions on prices and with foresight as to the effect of ED on the pricing game. It is relatively simple to adjust wholesale prices through time in response to cost changes or, as in the case of *Standard Fashion*, to demand conditions for various dress patterns subject to the vagaries of fashion demand; it is relatively complex to adjust a marketing strategy that alters either universally or selectively a decision to prevent retailers from carrying substitute products. Hence the assumption of a commitment to ED decisions at the time of pricing decisions.

Finally, we take the existence of both products as given, following both Bork and Comanor and Frech. The role of exclusivity clauses in protecting quasi-rents flowing from the development of new products against free riding by rivals, and in protecting specific asset values from opportunistic behavior (Howard Marvel, 1982) is a separate and recognized theme. This model as well as Bork and CF's analyses shows that an explanation of ED does not require free riding. Ignoring the alternative roles for ED does, however, bias our welfare analysis against the possibility that ED is welfare enhancing.

Let $C_1, C_2; W_1, W_2; P_1, P_2$ denote respectively the per unit manufacturing costs, and the wholesale and retail prices. $q_1(P_1, P_2)$, $q_2(P_1, P_2)$ define the retail demand functions of the two goods. $[W_i$, ED], $[W_i$, no ED] denote contract offers at wholesale price W_i with ED and without ED by manufacturer i. $R(W_1, W_2)$ is the retail profit when both products are bought at W_1, W_2: $R(W_1, W_2) \equiv \max(P_1 - W_1)q_1(P_1,P_2) + (P_2 - W_2)q_2(P_1,P_2)$. $\hat{R}^i(W_i)$ is the retail profit at wholesale price W_i when only the ith good is carried by the retailer. Finally, $\pi_i(W_1, W_2)$ is the profit (per retailer) accruing to manufacturer i when both products are carried. $\hat{\pi}_i(W_i)$ is manufacturer i's profit at W_i when only i's good is carried, that is, when the contract $[W_i$, ED] is accepted by the retailer.

The specific questions we address are: When does the equilibrium in contract offers by manufacturers include ED? What is the impact of prohibiting ED on market price, on profits, and on total welfare? The game is solved recursively. When neither firm has invoked ED, the subsequent price game results in the simple Bertrand duopoly prices (W_1^*, W_2^*); the payoffs to the firms are $\pi_1(W_1^*, W_2^*)$ and $\pi_2(W_1^*, W_2^*)$. When at least one of the two firms invokes ED, however, the retailer must choose between them, given their wholesale price offers. The retailer chooses the firm offering to the retailer the higher profits, $\hat{R}^i(W_i)$. If we index the two producers to satisfy $\hat{R}(C_1) \geq \hat{R}^2(C_2)$, the Nash equilibrium of the price game, after a decision to invoke ED, involves firm 1 offering a limit price $\hat{W}_1$ defined by $\hat{R}^1(\hat{W}_1) \equiv \hat{R}^2(C_2)$ and firm 2 offering a wholesale price equal to C_2. Thus the firm yielding the largest retail profits at a zero-wholesale markup (the "dominant" firm, labeled 1) captures the market in the price sub-game equilibrium subsequent to an ED offer; only at the pair of offers $(\hat{W}_1, C_2)$ has neither firm the incentive to change wholesale price offers. The structure and payoffs of the game are summarized in Figure 1.

Firm 2 obviously has no incentive to impose ED. ED is therefore observed in equilibrium if and only if $\hat{\pi}_1(\hat{W}_1) > \pi_1(W_1^*, W_2^*)$. The dominant firm's wholesale price falls with ED, that is, $\hat{W}_1 \leq W_1^*$, if and only if $\hat{R}^1(W_1^*) \leq R^2(C_2)$. This possibility is consistent with ED in equilibrium.

ED is profitable in spite of a decrease in the dominant firm's wholesale price when $\hat{R}^1(W_1^*)$ is only slightly less than $R^2(C_2)$. In this case firm 1 invokes ED, knowing that it need drop its wholesale price only slightly to ensure that it captures the market; the jump in demand when its rival is preempted more than compensates for the decrease in wholesale price. The retailer, in this case, is necessarily worse off. Of course, a sufficient condition for ED to be profitable is that it increases the price of the dominant firm, that

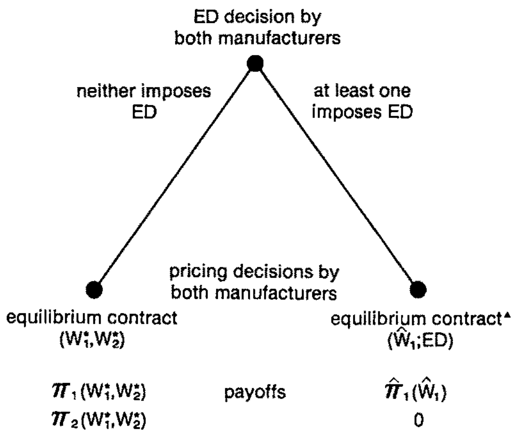


FIGURE 1. SUMMARY OF THE ED GAME: † FIRMS ARE ASSIGNED LABELS (1,2) TO SATISFY $R_1(C_1) \geq R_2(C_2)$; ▲ IN THE EQUILIBRIUM CONTRACT WITH ED, $R_1(\hat{W}_1) \equiv R_2(C_2)$

is, that $\hat{R}^1(W_1^*) \geq \hat{R}^2(C_2)$. The firm that finds ED potentially profitable is the firm that can afford the larger bribes to the retailers. Should the competitor attempt to match the rent transfers of the successful firm that still finds exclusive dealing worthwhile, the competitor would be bankrupt. To observe ED in this model, therefore, firms must be asymmetrical. If firms are "just asymmetrical enough" for ED to pay, then the wholesale price will drop with ED.

An examination of the retailer's first-order conditions reveals that the retail price may rise with ED even if the dominant firm's wholesale price falls. But with a sufficient drop in the wholesale price of the dominant firm, the retail price must drop. Welfare, consumers' surplus, is affected by ED in two ways. First, the selection of products in the market is reduced—a negative impact. Second, the retail price of the successful firm's product may rise or fall. If the retail price rises (or falls by a small amount), then consumers' surplus is unambiguously reduced. Welfare falls since the increase in profits to the dominant firm cannot offset the total decrease in the other three components of total surplus: consumers' surplus, retail profits, and the rival's profits—there is an associated deadweight loss. The retail price of good 1 will rise and welfare will fall with ED when the demand for the products is very asymmetric. Therefore, ED reduces welfare, in this model, when the private incentive for the restraint is greatest. Bork's contention that ED can only improve efficiency is incorrect.

The question remaining is whether ED can improve welfare: Can the price under ED drop so far as to overwhelm the negative effect on welfare of the reduction in product variety? If so, then the discipline of potential competition in the market is more powerful than the discipline of actual competition: Apparent (*ex post*) monopolies may be more efficient than (*ex post*) competition. An account of this issue requires more structure.

II. Welfare Pseudo-Empirics

Consider a linear demand example. After normalization, linear retail demand curves can be parameterized as $q_1 = \max[0, 1 - P^1 + c.\min(P^2, (1 + cP^1)/b)]$; $q_2 = \max[0, 1 - bP^2 + c.\min(P^1, 1 - cP^2)]$ with $c^2 > b$. When $q_1, q_2 > 0$, these demand curves become $q_1 = 1 - P^1 + cP^2$ and $q_2 = 1 - bP^2 + cP^1$. (The terms $1 - cP^2$ and $(1 + cP^1)/b$ are the "choke" prices in the two markets, respectively; they enter the demand curve definitions because as the price in one market is raised above its choke-off level, the demand in the other market should be unaffected.) The restriction of parameters to satisfy $c^2 > b$ guarantees that maximum profits are finite.

To focus exclusively on demand differences, we set variable retail costs at zero and measure retail profits gross of any fixed costs. The following calculations were made for the range of parameter values $10 \geq b > 1$ (varying degrees of product dominance) and $1 > c \geq 0$ (varying degrees of product substitutability): (*i*) Bertrand-Nash (no-ED) equilibrium: $\pi_i(W_1^*, W_2^*)$, $R(W_1^*, W_2^*)$, $CS(P_1^*, P_2^*)$; (*ii*) ED equilibrium: $\hat{\pi}_1(\hat{W}_1)$, $\hat{R}^1(\hat{W}_1)$, $C\hat{S}(\hat{P}_1)$.

Comparisons on the private and social incentives for ED in our sequential decision structure require the calculation of the ratio of manufacturer 1's profits (denoted by *PI*) and the ratio of the sum of consumers' plus producers' surpluses (*SI*) under the two re-

gimes. A private incentive exists for invoking ED if $PI(b,c) \geq 1$; a social incentive exists if $SI(b,c) \geq 1$. Figure 2 illustrates the contours for which $PI(b,c)=1$ and $SI(b,c)=1$. An understanding of the private and social incentives for ED emerges from Figure 2 if we fix one parameter and examine these incentives as the second parameter changes. First, fix c, the degree of substitutability, say at .8. ED is profitable for firm 1, provided that its competitor's market is sufficiently small (b sufficiently large). The rationale lies in the size of the implicit bribe to the retailer (in the form of a reduction in the wholesale price) required for exclusivity. If b is very close to 1 (the market nearly symmetric), then the bribe to the retailer necessary to meet firm 2's best ED offer is so large that the Bertrand profits dominate ED profits for firm 1. Firm 1's equilibrium wholesale price falls far enough with ED that the prospect of capturing the entire market is insufficient to render the restraint profitable. As b increases from 1, ED eventually involves a wholesale price decrease for firm 1 sufficiently small that it is overwhelmed by the jump in demand as the competitor is preempted: ED is profitable. As b increases further, a point is reached where firm 1 can guarantee itself the market by imposing ED without its wholesale price declining below the Bertrand level. Beyond this point (to the right of the dotted line in Figure 2), ED involves a price increase for firm 1. Finally, for b sufficiently large (b higher than 11.4 for $c=.8$), ED by firm 1 "blockades" entry by firm 1.

Now consider the social incentives for ED. Again fix c (say at .8). For low levels of b, welfare increases under ED because the reduction in wholesale price more than offsets the decrease in product variety. As b increases, the bribes to capture the retailer fall and eventually the price reduction fails to compensate for the choice-set restriction. A similar interpretation follows from holding b constant and changing c.[3]

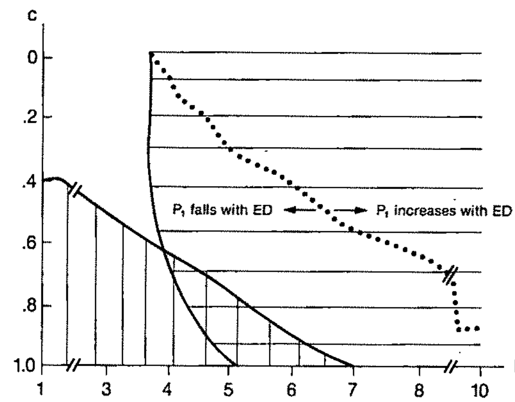


FIGURE 2. PRIVATE AND SOCIAL INCENTIVES FOR ED. HORIZONTAL LINES INDICATE THE REGION WHERE π_1 INCREASES WITH ED, A PRIVATE INCENTIVE FOR ED; VERTICAL LINES INDICATE THE REGION WHERE SURPLUS INCREASES WITH ED, A SOCIAL INCENTIVE. PARAMETERS b AND c ARE BOUNDED APPROPRIATELY (SEE TEXT)

The sets of b and c, where ED increases profit and where it increases welfare, intersect but do not coincide. The profitable use of ED does not necessarily increase welfare as Bork claims. But the effect of potential competition on prices may be strong enough that welfare is increased with ED but weak enough that the dominant firm is not dissuaded from invoking the restraint. The use of ED to preempt a rival—allegedly anticompetitive in Comanor and Frech's terminology—may be welfare increasing.

III. Conclusion

Modern industrial organization theory emphasizes the role of potential competition in disciplining markets, in contrast to the traditional focus on "actual" competition. We show in this paper that a public policy decision to allow exclusive dealing may enhance potential competition in wholesale markets, at the expense of a reduction in actual competition.

[3]In our analysis, we have identical variable costs for both firms. If this were not the case, there is an additional welfare effect. Any increase in the rival's variable cost relative to the dominant firm's increases the private incentives for exclusive dealing; welfare would be enhanced as production is shifted toward the more efficient firm.

Our analysis of the net competitive impact of exclusive dealing supports a middle ground between the positions of Bork and Comanor-Frech. CF argue that ED is anticompetitive when it excludes rivals, and by implication that it should then be disallowed in the absence of mitigating free-rider effects. We show that under ED, potential competition—competition for retail channels—may drive down the retail price,[4] even to a level where welfare increases with the restraint.

In a similar setting, Bork asserts that ED is invariably welfare improving. We argue that Bork's assertion rests upon an unrealistic assumption of retailer monopsony power. Retailer monopsony power does not follow from local retailer monopoly power.

Our model thus speaks against a per se approach to exclusive dealing. The rule suggested by Thomas Krattenmaker and Steven Salop (1986), that an exclusionary practice should be prohibited when it allows a firm to raise its price, is supported as a conservative rule of reason or *ex post* test of legality. By this test, exclusive dealing in the *Standard Fashion* case would have been allowed since it was initiated in exchange for a 50 percent discount on the prices of dress patterns.

The main positive result emerging from our model is a prediction of when exclusive dealing is likely to be observed. In this regard, we argue that a necessary condition for the profitability of ED, beyond product differentiation, is an *asymmetry* of the demand for the products. Symmetric product differentiation, even if substantial, is inconsistent with the observation of ED when manufacturers have foresight as to the effect of ED arrangements on price competition.

In our model, as in Bork and CF, both products exist before the ED game begins. We therefore ignore any use of ED in protecting the quasi-rent stream necessary to warrant product and retail development investment by either firm. We do not analyze the use of ED in franchise contracts with possibly two-part franchise fees, where bribes to retailers to carry a product exclusively need not come through lowered wholesale prices.

[4]A closely related idea is Harold Demsetz's (1968) alternative to the regulation of public utilities. Demsetz suggests auctioning the right to serve a natural monopoly market. The bids take the form of prices which the candidate firms would charge in the market with the lowest price winning. The Demsetz scheme is not actually observed, presumably because its implementation has incentive and commitment problems associated with the need for long-lived specific assets in most natural monopolies. For vertical restraint policy, our model shows that allowing ED (which is a feasible policy) corresponds exactly to the creation of Demsetz's competition in the market.

REFERENCES

Bork, Robert H., *Antitrust Paradox*, New York: Basic Books, 1978.

Comanor, William S. and Frech, H. E., "The Competitive Effects of Vertical Agreements," *American Economic Review*, June 1985, *75*, 539–46.

Demsetz, Harold, "Why Regulate Utilities?," *Journal of Law and Economics*, April 1968, *11*, 55–65.

Krattenmaker, Thomas G. and Salop, Steven, "Anticompetitive Exclusion: Raising Rivals' Costs to Achieve Power Over Price," *Yale Law Journal*, December 1986, *96*, 209–93.

Marvel, Howard, "Exclusive Dealing," *Journal of Law and Economics*, April 1982, *25*, 1–25.

Schwartz, Marius, "The Competitive Effects of Vertical Agreements: Comment," *American Economic Review*, December 1987, *77*, 1063–68.

Standard Fashion Co. v. Magrane-Houston Co., 258 U.S. 346, 42 S.Ct. 360, 66 L.Ed. 653 (1922).

The Competitive Effects of Vertical Agreements: Comment

By Marius Schwartz*

In a recent paper, Comanor and Frech (1985, "CF") claim that a manufacturer enjoying a product differentiation advantage could impose exclusive dealing on distributors while also raising its wholesale price. The two consumer groups in the model would then pay higher prices than under nonexclusive dealing. CF also suggest that exclusive dealing is more likely to emerge when product differentiation is relatively strong.

CF's analysis does not carefully incorporate the constraints imposed by dealer rationality. Doing so within CF's model, I reach opposite conclusions. Exclusive dealing will be accepted only if the manufacturer reduces its wholesale price to dealers. Relative to nonexclusive dealing, price rises to one group of consumers but falls to the other. Moreover, exclusive dealing is more likely to emerge when product differentiation is relatively weak.

I. Comanor and Frech's Model

CF consider a dominant manufacturer M facing rival manufacturers E. There are two groups of consumers, A and B. Group B views the product as identical, while group A is willing to pay a constant premium of α for M's product. Price discrimination by M between the two groups is ruled out. Manufacturers M and E both have constant cost c.

There are two dealer groups: low cost L and high cost H, and the number of L dealers is fixed. The distribution margins of L dealers and H dealers are assumed constant throughout at γ and δ, respectively. Distribution margin is the difference between a dealer's price to consumers (retail price) and the manufacturer's price to the dealer (wholesale price). The assumption that distribution margins are fixed—through unspecified "monopolistic competition" among dealers—is unappealing, but is not critical to CF's argument. For instance, H dealers can be viewed as perfectly competitive with constant unit-cost δ. The key point is that a manufacturer incurs lower distribution cost by selling through L dealers. This will hold provided there is "enough" competition among L dealers to keep their margin (which includes both cost and any profit) below H dealers' cost δ. In order to pin down retail and wholesale prices, CF assume that L dealers' margin is constant throughout at $\gamma < \delta$. I retain this assumption to stay within their model.

Under nonexclusive dealing, all manufacturers sell through L dealers. Assuming that manufacturers E are perfectly competitive, they set wholesale price equal to cost c. For convenience, I measure all prices as markups over the common cost c, hence E's wholesale price is $P_E = 0$. The retail price, denoted by a prime throughout, is then $P_E' = \gamma$. Since M would have to set wholesale price at zero to capture consumers B, its optimal strategy is to abandon them and extract the full brand premium α from consumers A. Purely for convenience, I adopt throughout the tie-breaking rule that whenever a consumer is indifferent between the two products, he purchases from M. Thus, the nonexclusive dealing equilibrium has

$$(1) \qquad P_E = 0, \quad P_E' = \gamma;$$

$$P_M = \alpha, \quad P_M' = \gamma + \alpha.$$

If M obtains exclusivity with L dealers, P_E' rises from γ to δ, since only H dealers would carry E's product. M can now pursue one of

*Georgetown University and Economic Analysis Group, Antitrust Division, United States Department of Justice. The views expressed here do not necessarily reflect those of the United States Department of Justice. I wish to thank Curt Jernigan, John Kwoka, Marilyn Simon, two referees of this *Review*, and especially Tim Brennan, Ted Frech, David Malueg, and Albert K. Smiley. Unfortunately, all errors are mine.

TABLE 1—OUTCOMES IN VARIOUS REGIMES

Dealing	M's Strategy	M Serves	E Serves	P_E'	P_M'	$\Delta P_E'$	$\Delta P_M' = \Delta P_M$
Nonexclusive	I	A	B	γ	$\gamma+\alpha$	—	–
Exclusive	I	A	B	δ	$\delta+\alpha$	$\delta-\gamma$	$\delta-\gamma$
Exclusive	II	A, B	–	δ	δ	$\delta-\gamma$	$\delta-\gamma-\alpha$

two strategies:

(2) I. high-price strategy:
set P_M to capture only consumers A.

II. low-price strategy:
set P_M to capture consumers A and B.

Thus, the exclusive-dealing prices are[1]

$$(3)\qquad P_E = 0;\quad P_E' = \delta,$$

$$P_M = \begin{cases} \delta-\gamma+\alpha & \text{if I,} \\ \delta-\gamma & \text{if II.} \end{cases}$$

$$P_M' = \begin{cases} \delta+\alpha & \text{if I,} \\ \delta & \text{if II.} \end{cases}$$

Table 1 summarizes the three regimes: nonexclusive dealing, and exclusive dealing under strategy I or II. The changes $\Delta P_E'$ and $\Delta P_M'$ are measured relative to nonexclusive dealing. The striking feature of the table is that by moving to exclusive dealing M might be able to raise wholesale (and retail) price, whether pursuing strategy I or II. Under strategy I, capturing only group A, $\Delta P_M \equiv \Delta P_M' = \delta - \gamma > 0$. Under strategy II, capturing both consumer groups, $\Delta P_M \equiv \Delta P_M' = \delta - \gamma - \alpha$, which is positive if $\delta - \gamma > \alpha$.

The table, which summarizes CF's Section III, incorporates only the constraint on pricing imposed by the availability of product E from H dealers. There is a second constraint, however, implied by the need to obtain acceptance of exclusivity by L dealers. CF discuss this only later in their Section IV. It is difficult to tell how their discussion modifies the conclusions of the table, since in their Section IV a key variable λ is defined and used inconsistently (see Marius Schwartz, 1985, Appendix). I therefore begin by reexamining their discussion, while also making explicit the nature of M's offer to L dealers.

II. Wholesale Price Under an All-or-None Offer

CF's Section IV implicitly assumes that M offers exclusivity at P_M to L dealers and makes the offer all-or-none, that is, contingent on all dealers accepting. (I will show in Section III that the offer must be all-or-none in order to avoid the usual holdout incentive.) Each L dealer expects that, were it to reject M's offer, M would sell only through H dealers if it could do so at some price $\hat{P}_M \geq 0$ (recall that all prices are measured above cost).[2] If L dealers accept exclusivity with M at P_M, retail prices will be $P_M' = P_M + \gamma$, $P_E' = \delta$. If they reject M's offer and purchase exclusively from E at 0, prices will be $P_M' = \hat{P}_M + \delta$, $P_E' = \gamma$. Let $Q_j^i(P_M', P_E')$ denote demand by consumer group j for product i when both products are available at retail prices P_M' and P_E', $i = E, M$, and $j = A, B$. L dealers will therefore accept exclusivity with M at P_M only if

$$(4)\qquad \frac{\gamma}{n}\left[Q_A^M(P_M+\gamma,\delta)+Q_B^M(P_M+\gamma,\delta)\right]$$

$$\geq \frac{\gamma}{n}\left[Q_A^E(\hat{P}_M+\delta,\gamma)+Q_B^E(\hat{P}_M+\delta,\gamma)\right],$$

[1] These prices are upper bounds on M's retail price, imposed by the availability of product E from H dealers. If demand were sufficiently elastic, M's optimal prices would be lower. These upper bounds therefore can overstate the welfare loss from exclusive dealing, but to simplify the discussion I assume that the prices in (3) are more profitable than lower ones.

[2] I am grateful to Ted Frech for clarifying this in private communication.

where n is the number of L dealers and γ is their constant margin.[3]

Let $Q_j^i(P_i')$ denote the demand by group j for product i when only product i is available. Since the products are perfect substitutes for group A when $P_M' = P_E' + \alpha$ and perfect substitutes for group B when $P_M' = P_E'$, demands satisfy

$$(5)\quad [Q_A^M(P_M', P_E'), Q_A^E(P_M', P_E')]$$
$$= \begin{cases} [Q_A^M(P_M'), 0] & \text{if } P_M' \leq P_E' + \alpha \\ [0, Q_A^E(P_E')] & \text{if } P_M' > P_E' + \alpha. \end{cases}$$
$$[Q_B^E(P_M', P_E'), Q_B^E(P_M', P_E')]$$
$$= \begin{cases} [Q_B^M(P_M'), 0] & \text{if } P_M' \leq P_E' \\ [0, Q_B^E(P_E')] & \text{if } P_M' > P_E'. \end{cases}$$

Consider first what CF call *weak brand preference*, $\alpha < \delta - \gamma$. That is, the brand premium is less than the margin differential between dealers. L dealers will now obviously reject exclusivity at any $P_M \geq \alpha$. By purchasing from E at price 0, they can offer E's product at $P_E' = \gamma$ and capture both consumer groups.[4] Since group A's demand curve for product E equals its demand for product M shifted down by α, while group B's demand curves for E and M are identical, offering E at $P_E' = \gamma$ yields L dealers higher sales (and profit) than had they accepted M's offer of $P_M \geq \alpha$ and set $P_M' = P_M + \gamma$. Thus, under weak brand preference, M cannot achieve exclusivity unless it reduces wholesale price below its nonexclusive-dealing level of α. This contrasts with the impression conveyed by CF, as summarized in Table 1.

Under *strong brand preference*, $\alpha > \delta - \gamma$, if M wishes to capture both consumer groups (strategy II) it must still set $P_M < \alpha$. To see this, note that for $P_M = \alpha$ (5) implies $Q_B^M(P_M', P_E') = 0$, since $P_M' = \alpha + \gamma$, $P_E' = \delta$ (E sells via H dealers) and strong brand preference means $\alpha + \gamma > \delta$ or $P_M' > P_E'$. In this case, it is product E offered via H dealers that forces M to reduce wholesale price if it wishes to capture groups A and B. However, if M wishes to capture only group A (strategy I), it might be able to achieve exclusivity while raising its wholesale price above α. To see this, note that for any $P_M = \alpha + \varepsilon$, $0 < \varepsilon < \delta - \gamma$, equations (5) reduce condition (4) to $Q_A^M(\alpha + \varepsilon + \gamma) \geq Q_B^E(\gamma)$, which can be met if group A is larger than B. Thus, under strong brand preference, it seems that M could obtain exclusive dealing at a wholesale price higher than it charged under nonexclusive dealing, while retaining consumers A and thereby increasing its profit.

III. Plausibility of an All-or-None Offer

The possibility of P_M rising above its nonexclusive-dealing level of α is intriguing, since α is the maximum premium that consumers A place on M's product. It is worth scrutinizing the assumptions underlying this possibility.

First, observe that M's exclusivity offer must be all-or-none, otherwise every L dealer will want to be the holdout. If an individual L dealer expects that it can unilaterally reject M's exclusivity without affecting M's price to the remaining L dealers, it will stay with M only if such *unilateral* defection to E is unprofitable:

$$(6)\quad \frac{\gamma}{n}[Q_A^M(P_M + \gamma, \delta) + Q_B^M(P_M + \gamma, \delta)]$$
$$\geq P_E'[Q_A^E(P_M + \gamma, P_E') + Q_B^E(P_M + \gamma, P_E')],$$

where P_E' is the switching dealer's retail price that equals its margin (recall that $P_E = 0$). Given any $P_M > \alpha$, there are values of P_E'

[3]This condition is necessary but not sufficient, since L dealers may believe that if they reject M's offer of P_M, M might offer a lower price rather than turning exclusively to H dealers at $\hat{P}_M$. Condition (4) thus presumes that L dealers view M's choice as being solely between dropping out of the market and selling to H dealers. See also Section V.

[4]The lowest retail price for M's product would then be $P_M' = \delta$, achieved by setting $P_M = 0$ to H dealers. Since $\delta - \gamma > \alpha$ in this case, (5) shows that consumers A, as well as B, would purchase only product E.

that make this inequality fail, for example, $P_E' = \gamma$ (use (5)). Note well that this incentive for unilateral defection must be addressed, because CF's argument implicitly requires that L dealers not be a collusive group. For if they were collusive, they would have raised their margin from γ to the feasible maximum δ, all manufacturers would be indifferent between them and H dealers, and exclusive dealing would be irrelevant.

Next, observe that preventing unilateral defection requires M to threaten that it would respond to defection by reducing its price to rival dealers below some level $\bar{P}_M$, enough to make the defector's profit lower than what it would earn by remaining exclusive with M (left-hand side of (6)). In general, M's best response to unilateral defection will *not* be to drop price to $\bar{P}_M$ to rival dealers, so a precommitment ability is necessary to prevent defection. This can be seen as follows.

The level of $\bar{P}_M$ depends on whether M will sell through the H dealers, as assumed in Section II, or through the nondefecting L dealers. The latter allows a higher $\bar{P}_M$ (higher by the cost differential, $\delta - \gamma$) so suppose that is how M would respond to defection. If making defection unprofitable requires denying the defector both consumer groups, A and B, then $\bar{P}_M = 0$, since a defecting L dealer purchasing at $P_E = 0$ could undercut loyal L dealers purchasing at any $P_M > 0$ and thus capture group B.[5] But $P_M = 0$ obviously is not M's best response (for example, any $P_M \leq \alpha$ will retain all consumers A). If making defection unprofitable only requires denying the defector consumers A, then $\bar{P}_M = \alpha$. This would constitute M's best response if setting any $P_M > \alpha$ would mean losing all consumers A to the defector. But realistically, a single L dealer will not have sufficient capacity to serve the entire A group so it will expect that in response to its single defection M probably will maintain P_M somewhat above α to remaining L dealers.

In short, M's best response to unilateral defection generally would not render defection unprofitable, so M must be committed to executing its threat of reducing price. But then it is such commitment ability, not M's product differentiation, that allows M to impose exclusivity while raising price. Any E manufacturer could similarly obtain exclusivity with L dealers, enabling it to raise its wholesale price from zero to $\delta - \gamma$ while selling to group B, if it could commit to punish any L dealer that refused to cut off other E manufacturers.[6] Similarly, any one of the L dealers could attain dealer collusion, thereby raising dealer margin from γ to δ, if it could commit to punish any defectors. No convincing reason has been offered for why M has an advantage in making punishment commitments.

Finally, suppose that M *could* credibly commit to charge $P_M \leq \alpha$ to other dealers (whether H dealers or loyal L dealers) if even one L dealer rejects exclusivity at $P_M > \alpha$. It is *doubtful* that M *would* choose to make such an all-or-none offer. Making all-or-none offers is risky in general, because of the possibility that one "crazy" may not go along. (This could be overcome if the potential "crazies" could be identified and exempted from the all-or-none offer, which may or may not be feasible.) In this context, an all-or-none offer is particularly dangerous because it creates strong incentives for various parties to bribe an individual L dealer to refuse. Such refusal would trigger a lower price by M in order to punish the holdout. The beneficiaries from M's price reduction —whether they be the other L dealers or all H dealers—will therefore have a powerful

[5]It is likely that defection will be profitable if the defector can compete for group B, because it will be able to capture all this group. Unless group B is much smaller than group A, this will outweight the *fraction* of group A that the defector loses by abandoning exclusivity with M.

[6]Administering the punishment would require E's reducing price below cost (setting $P_E < 0$), but it is not obvious why this threat is any less credible than M's. A native antitrust rule that interpreted predatory pricing as pricing below cost might preclude this, while allowing M's price reduction. But it is doubtful that courts interpret price-below-cost as either necessary or sufficient for inferring predation and, even if they did, it is likely that E could find other feasible ways to punish a refusing L dealer.

incentive to raise a bribe sufficient to induce holdout by a single L dealer.[7] Recognizing this, M would prefer to deal nonexclusively at $P_M = \alpha$, where L dealers have no holdout incentive.

IV. Exclusive Dealing and Brand Preference

Consider now CF's claim that exclusive dealing is more likely to emerge when brand preference is strong rather than weak. CF's claim is based on the notion that under strong brand preference M might be able to achieve exclusivity while raising P_M above α (see their Section II). We have seen that this is highly implausible, as it requires M credibly threatening to punish any defector—even though implementing the punishment is costly for M and even though there is a good chance that the punishment will have to be implemented.

Given the need to set $P_M < \alpha$ to obtain exclusivity, M will offer exclusivity only if it plans to follow strategy II, capturing both consumers A and B. (If only consumers A were captured and $P_M < \alpha$, M obviously would do better by remaining nonexclusive at $P_M = \alpha$.) Since strategy II is more likely to increase M's profit relative to nonexclusive dealing when brand preference is weak rather than strong, exclusive dealing is more likely when brand preference is weak.

To see this, let P_M^* be M's optimal price under exclusive dealing. P_M^* must satisfy two constraints: (a) $P_M^* < \alpha$, for acceptance of exclusivity by L dealers; and (b) $P_M^* + \gamma \leq \delta$, so that customers B are not lost to manufacturers E selling through H dealers. Constraint (b) implies $P_M^* \leq \delta - \gamma$, which requires $P_M^* < \alpha$ only if $\delta - \gamma < \alpha$, the case of strong brand preference. Thus, constraint (b) is certainly not binding under weak brand preference but may be binding under strong. The decrease in P_M from its nonexclusive-dealing level of α, therefore, will be at least as large when brand preference is strong. Intuitively, strong brand preference means that the margin that M can extract from consumers A under nonexclusive dealing (α) exceeds the maximum margin it could extract from consumers B under exclusive dealing with strategy II ($\delta - \gamma$). This tends to make it more attractive for M to stick with only customers A and forego exclusive dealing.

[7]It is true that this requires some cooperation among the prospective beneficiaries, but the requisite cooperation will be "small" since the bribe that must be raised will be small compared to the total gain to the group. It also is true that H dealers have been assumed to be perfectly competitive and thus always earning zero profit, but this assumption was merely for convenience and could easily be relaxed.

V. Concluding Remarks

The conclusions that exclusive dealing (i) requires lowering the wholesale price; and (ii) is more likely when product differentiation is relatively weak are sensitive to the way CF represent product differentiation—as the size of a constant premium that the manufacturer can fully extract through its wholesale price. Matthewson and Winter ("MW," 1987) obtain different results by considering two manufacturers, 1 and 2, selling *imperfect*-substitute products (with each product's demand a continuous function of both prices).

Selling to a local-monopolist dealer, under nonexclusive dealing the manufacturers charge the Bertrand-equilibrium wholesale prices (W_1^*, W_2^*). Under exclusive dealing, the dominant manufacturer 1 sets a wholesale price $\hat{W}_1$ that yields the dealer equal profit as buying exclusively from 2 at 2's cost (the lowest wholesale price 2 will offer). MW show that 1's profit can be higher at $\hat{W}_1$ than at (W_1^*, W_2^*) and, more surprisingly, that $\hat{W}_1 > W_1^*$ is possible. That is, exclusivity might rationally be accepted even if manufacturer 1 raises the wholesale price. Their examples also indicate that exclusive dealing is more likely when 1's product differentiation advantage is relatively strong.

The imperfect substitutability between the products is the key to MW's findings. If the products were perfect substitutes (either at the same price or at a constant differential α, as in CF), the equilibrium would always have 1 as a monopolist charging a price equal to 2's cost (plus any brand premium)—whether or not 1 required exclusivity. The

different representation of product differentiation also explains why MW find exclusive dealing more likely when product differentiation is strong, while CF's model implies the reverse as I have shown.

Both CF and MW grant the dominant manufacturer the ability to refuse to deal nonexclusively. This entails some commitment power, since if the dealer(s) held out the manufacturer would be better off accepting nonexclusive dealing than the alternative (of dropping out in MW and of turning to *H* dealers in CF). But CF require considerably more commitment ability, since their manufacturer necessarily faces multiple dealers in any local market and must therefore commit to an all-or-none offer that effectively mimics what collusion among the dealers would have achieved. MW's demonstration of price-raising exclusive dealing uses a different and sometimes overlooked principle: that, except under very special demands, full exploitation of a product differentiation advantage requires multiple instruments rather than just the wholesale price.

REFERENCES

Comanor, William S. and Frech, H. E. III., "The Competitive Effects of Vertical Agreements?" *American Economic Review*, June 1985, *75*, 539–46.

Matthewson, F. B. and Winter, R. A., "The Competitive Effects of Vertical Agreements: Comment," *American Economic Review*, December 1987, *77*, 1057–62.

Schwartz, Marius, "The Competitive Effects of Vertical Agreements: Comment," Economic Policy Office Discussion Paper EPO 85-9, Antitrust Division, U.S. Department of Justice: Washington, DC, August 1985.

The Competitive Effects of Vertical Agreements: Reply

By WILLIAM S. COMANOR AND H. E. FRECH III*

In our earlier paper (1985), we constructed a simple model of market relationships between manufacturers and distributors, which demonstrated that exclusive dealing arrangements can be used to exclude entrants and thereby have anticompetitive effects. For this reason, in antitrust cases involving exclusive dealing arrangements, we favored an antitrust standard of the "rule of reason" rather than an irrebuttable presumption that all such arrangements are legal per se.[1] In their comments, Frank Mathewson and Ralph Winter (1987) and Marius Schwartz (1987) have criticized and extended our analysis but not altered our essential conclusions.

These writers construct models in which exclusive dealing arrangements can be profitable, although their models differ from each other's as well as from our own. Mathewson and Winter assume a different structure of substitution between the products of a dominant manufacturer and an entrant than we had posited. Schwartz, on the other hand, differs in the type of commitment that a dominant manufacturer can make to induce distributors to accept exclusive dealing. Still, both confirm our basic findings.

Both comments acknowledge that firms may gain from the imposition of exclusive dealing arrangements even while consumers are harmed. They also suggest that exclusive dealing arrangements can sometimes lead to improved consumer welfare, a finding consistent with our own result that exclusive dealing can lead to lower prices for some consumers (1985, p. 541). Our differences arise not over the possibility of such results, but over the likelihood that such circumstances may in fact arise.

Mathewson and Winter write that exclusive dealing can lead to improved welfare because the potential competition from new entrants may be more effective at constraining prices than actual competition. This means that prices can fall to at least some consumers because the imposition of exclusive dealing alters the margins on which both manufacturers and distributors compete. If the new demand conditions are more elastic, prices can decline.

These results are more easily seen in our model. Without exclusive dealing, a manufacturer profits most from charging a high price and selling only to those consumers who prefer his brand to others. With exclusive dealing arrangements, he is more likely to compete for those consumers who do not have strong brand preferences. As a result, prices can decline to strongly brand-conscious consumers. Depending on the magnitude of the price changes, and the relative number of brand-conscious consumers, economic welfare can improve or decline.

I. Mathewson and Winter's Model

Mathewson and Winter (MW) assume that "two manufacturers sell products to a large number of retailers" (p. 1057) or distributors. Furthermore, one of these firms is "dominant" in that he "captures the market" (p. 1057) if he imposes exclusive dealing and still makes positive profits. Therefore, the dominant firm's profit function cannot be duplicated by either a second producer or an entrant to the market. These conditions are similar to our own.

The structure of substitution between the products of the two producers distinguishes the MW model from ours. Their structure is more general and therefore has less economic content. It is consistent with any number of consumers who are either differ-

*Department of Economics, University of California, Santa Barbara, CA 93106. We are grateful for the helpful comments of M. Bruce Johnson, Clement Krouse, Steven Salop, and Perry Shapiro.

[1] That suggestion was made by Howard Marvel, 1982.

ent or alike. In contrast, we assume that some, but not all, consumers have a strong preference for the product of the original manufacturer, which leads to its dominant position.

In the MW model, the dominant manufacturer might be required to set a lower wholesale price to assure that his distributors accept exclusive dealing. While our earlier paper (1985) acknowledged this possibility (p. 543), it is hardly likely to be in the manufacturer's best interest. On the contrary, we suggested before that distributors will generally accept exclusive dealing even without lower prices where consumers show strong preferences for the manufacturer's product (pp. 542–43). The leading manufacturer's price cuts are designed more to attract price-sensitive consumers than to induce distributors to accept exclusive dealing.

Although the two studies formally give similar results, MW provide less indication of where exclusive dealing arrangements are likely to be found. Yet, we both agree that exclusive dealing is more likely where the product differentiation advantages of the leading firm are greater.

II. Schwartz's Model

Schwartz questions the way we model the relationship between the dominant manufacturer and his distributors that lead to the latter's acceptance of exclusive dealing. He misinterprets our paper, but even without that, there is a basic difference between us in the approach used to model strategic interactions.

Schwartz believes that, in our model, the dominant manufacturer must commit himself to stop dealing with all existing distributors and sell exclusively through higher-cost, alternate channels of distribution if any one distributor, in any local market, were to reject exclusive dealing. On the contrary, we assumed that local markets were sufficiently segmented so that no costly, universal boycott would be necessary. The manufacturer need only threaten to sell through alternate distribution channels in the same geographic market as that serviced by the distributor who rejected exclusive dealing.

If existing distributors were sufficiently competitive that exclusive dealing would unravel with the defection of a single distributor, the practice would be an unlikely outcome. Exclusive dealing for exclusionary purposes requires market power at the distribution stage.

Furthermore, the dominant manufacturer would be unlikely to stop all sales to existing distributors simply because one would not accept the exclusive dealing requirement. The narrowly rational response would be for the manufacturer to continue selling, on a nonexclusive basis, to the rebellious distributor. If this were known, however, distributors would have little incentive to accept exclusive dealing. Without a stronger response, the manufacturer must lower his price if he wishes to induce distributors to accept exclusive dealing. This, in fact, is Schwartz's model.

Schwartz requires that the manufacturer must be narrowly rational at all times, and thereby restricts outcomes to the perfect equilibria of game theory (Reinhard Selten, 1975; Duncan Luce and Howard Raiffa, 1957, pp. 97–102). He complains that our equilibria, as well as those of MW, are not perfect. We agree but believe that a wider class of equilibria is essential to understand oligopoly behavior. Furthermore, as Luce and Raiffa anticipated, the implications of the narrow rationality assumption have been falsified experimentally (for example, Lave, 1962). The narrow rationality/perfect equilibria assumption is also inconsistent with market behavior such as collision, or even exchange.[2]

There are additional reasons for rejecting the assumption of narrow rationality. Substantial advantages often accrue to an individual or firm in a bargaining context from being perceived as irrationally tough. Such advantages provide an incentive to make and

[2] Other behavior that conflicts with perfect equilibria includes the production of high-quality goods (Benjamin Klein and Keith Leffler, 1981); wartime cooperation (Robert Axelrod, 1984) and animal behavior (Axelrod, 1984; Robert Trivers, 1971; Jack Hirshleifer, 1977).

keep commitments that transcend narrow rationality.[3] David Kreps and Robert Wilson (1982) have shown that even a small probability of carrying out a threat may lead to its being perceived as credible, which can then influence the actions of others.

Firms and individuals can and do commit themselves to follow narrowly irrational paths, and therefore present others with a tougher reaction function than the simple accommodation suggested by Schwartz (Earl Thompson and Roger Faith, 1981; Thomas Schelling, 1960). The study of threats and commitments has long been a part of antitrust analysis (Lester Telser, 1966; Richard Posner, 1976, p. 186; and Oliver Williamson, 1985, pp. 376–77). To require, as Schwartz does, that firm conduct is always narrowly rational, is to condemn the economic analysis of strategic problems to an inaccurate view of market behavior.

We do not argue that any commitment imaginable can be made credible. Economists must rely on empirical and even historical observation of the interaction among firms, particularly in the same industry. Just like the choice between modeling price-setting or quantity-setting behavior, judgment and observation are essential. One must also examine why a particular firm can make a commitment, but not others. The analysis is necessarily asymmetric. Indeed, one of the advantages in being first in a market is the ability to make commitments.[4] The search for means by which particular commitments are made credible is worthwhile, but limiting oneself to commitments that are narrowly rational in repetitive games, as Schwartz recommends, is too constraining for most competitive problems among small numbers of firms.

Schwartz claims there is no reason why the entrant could not make a commitment similar to that made by the dominant manufacturer. Thus, he writes that "no convincing reason has been offered for why M (the dominant manufacturer) has an advantage in making punishment commitments." (p. 1066) But there is a good reason why the entrant could not make a similar commitment.

The fundamental asymmetry in our model is that some consumers view the dominant manufacturer's product as superior. But there is no such view of the entrant's product. Where the brand identification of the original manufacturer's product is substantial, a distributor faces reduced sales and profits if he cannot sell it. On the other hand, there are many potential entrants who are capable of producing the same product at the same prices, so that a distributor could simply replace one entrant's product with that of another. The asymmetry between the established manufacturer and the entrant is an important part of our model.

The model proposed by Schwartz rests on very different premises. Yet, we both find that exclusive dealing can sometimes promote higher prices and sometimes lower ones. His qualitative conclusion is consistent with our own.

III. Effects on Competition

Our earlier paper was limited to effects on prices and competition. This focus follows from the expressed goal of antitrust, which is increased competition. Yet, in recent years, the law has paid increasing attention to efficiency or consumer welfare. These additional concerns were introduced in the *Continental T.V. v. GTE Sylvania* (1977) decision, but have become evident in more recent decisions as well. Still, competitive effects and efficiency concerns are not treated symmetrically in antitrust actions or in economic theory. There is reason, therefore, to consider them separately.

Efficiency is often an available defense for actions that have anticompetitive results. Even where reduced competition can be demonstrated, it might still be tolerated where efficiency is advanced. An efficiency defense arises in the circumstances of a

[3]According to Robert Trivers (1971), this fact provides an evolutionary basis for the emotions of anger and indignation that can lead to narrowly irrational actions. It also explains the selection of business and political leaders, partly on character traits.

[4]See William S. Comanor and H. E. Frech III, 1984, p. 374.

specific case that is considered under the "rule of reason."

Mathewson and Winter examine efficiency considerations in Section II of their paper. Despite the title, that section simply provides an example. But the example rests on parameter estimates of a kind that would be difficult to discover in an actual case. The example provides little guidance for an actual efficiency defense.

IV. Concluding Comment

Despite the evident differences between our approach to exclusive dealing arrangements and those of both Mathewson and Winter and Schwartz, there are important similarities as well. Their results, as well as our own, suggest that circumstances may exist where the exclusive dealing arrangements have anticompetitive consequences. The implication of all three studies for antitrust enforcement is that these arrangements should be examined under the "rule of reason."

REFERENCES

Axelrod, Robert, *The Evolution of Cooperation*, New York: Basic Books, 1984.

Comanor, William S. and Frech, H. E. III, "Strategic Behavior and Antitrust Analysis," *American Economic Review*, May 1984, *74*, 373–76.

______ and ______, "The Competitive Effects of Vertical Agreements?," *American Economic Review*, June 1985, *75*, 539–46.

Hirshleifer, Jack, "Economics from a Biological Viewpoint," *Journal of Law and Economics*, April 1977, *20*, 1–52.

Klein, Benjamin and Leffler, Keith, "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, August 1981, *89*, 615–42.

Kreps, David P. and Wilson, Robert, "Reputation and Imperfect Information," *Journal of Economic Theory*, 1982, 245–52.

Lave, Lester B., "An Empirical Approach to the Prisoners' Dilemma Game," *Quarterly Journal of Economics*, August 1962, *76*, 424–36.

Luce, R. Duncan and Raiffa, Howard, *Games and Decisions*, New York: Wiley & Sons, 1957.

Marvel, Howard P., "Exclusive Dealing," *Journal of Law and Economics*, April 1982, *25*, 1–25.

Mathewson, G. F. and Winter, R. A. "The Competitive Effects of Vertical Agreements: Comment," *American Economic Review*, December 1987, *77*, 1057–62.

Posner, Richard A., *Antitrust Law: An Economic Perspective*, Chicago: University of Chicago Press, 1976.

Schelling, Thomas C., *The Strategy of Conflict*, London: Oxford University Press, 1960.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2nd ed., Chicago: Rand McNally, 1980.

Schwartz, Marius, "The Competitive Effects of Vertical Agreements: Comment," *American Economic Review*, December 1987, *77*, 1063–68.

Selten, Reinhard, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, August 1975, *4*, 25–55.

Telser, Lester G., "Cutthroat Competition and the Long Purse," *Journal of Law and Economics*, October 1966, *9*, 259–77.

Thompson, Earl A. and Faith, Roger I., "A Pure Theory of Strategic Behavior and Social Institutions," *American Economic Review*, June 1981, *71*, 366–80.

Trivers, Robert L., "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology*, March 1971, *46*, 35–58.

Williamson, Oliver E., *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

Continental T.V., Inc., v. GTE Sylvania, Inc., 433 U.S. 36, 1977.

ERRATUM

The Value of Federal Mineral Rights, Correction and Update

By MICHAEL J. BOSKIN AND MARC S. ROBINSON*

Michael Boskin, Marc Robinson, Terrence O'Reilly, and Praveen Kumar (BROK), using a methodology which they developed, estimated the value of federal oil and gas rights. In revising and updating the estimates for use in Boskin, Robinson, and Alan Huber (1987), we discovered a programming error underlying the 1981 benchmark estimate. In Table 1, we present a corrected series corresponding to Table 2 in BROK; the revised estimates are about one-third lower than those originally presented.

Table 1 also updates the BROK estimates. Prices of oil and gas have fallen; hence, the current value of federal mineral rights is lower than in 1981. Since the BROK methodology is designed to give a contemporaneous, rather than *ex post*, estimate, only the 1986 value reflects the 1986 price shock. The fall in prices, as well as bonus and royalty payments, caused the estimated value to drop from $521 billion in 1981 to $491 billion in 1985.[1] In real terms, the estimated value had dropped by 20 percent. The price collapse in 1986 caused an additional capital loss of $157 billion, or about two-thirds the size of the traditional budget deficit. The recent rebound in oil prices has erased more than half of the 1986 loss in federal oil rights.[2]

Though the magnitude of the fall may be overstated if oil and gas prices are expected to rebound, it is clear that the value of federal mineral rights has dropped. This reflects a true capital loss for the federal government, which should be taken into account in a comprehensive measure of national wealth.[3]

The conclusions of the original paper remain unchanged: the value of federal oil and gas rights is quantitatively significant relative to other assets and liabilities of the federal government and also fluctuates substantially as the result of price changes.

*Department of Economics, Stanford University, Stanford, CA 94305, and Operating Sciences Department, General Motors Research Laboratories, Warren, MI, 48090, respectively. We gratefully acknowledge the helpful comments and assistance given to us by Alan Huber and Donald Rosenthal.

[1]BROK's methodology is based on geologists' estimates of reserves, both proven and undiscovered (but economically recoverable at current prices and technology) on federal land in 1981. In recent years, dry holes in a number of promising areas may have reduced estimates of undiscovered reserves. In addition, the fall in prices makes some oil which was previously economic no longer worth recovering, though future price increases may restore the status quo. These revisions are not taken into account in Table 1.

[2]Spot prices for West Texas intermediate crude oil in July 1987 were 36 percent above average posted prices for 1986.

[3]See Boskin, Robinson, and Alan Huber (1987) for estimates of government tangible wealth and further discussion.

REFERENCES

Boskin, Michael J., Robinson, Marc S. and Huber, Alan, "Government Saving, Capital Formation and Wealth in the United States, 1947–1985," NBER Working Paper, 1987, to appear in Robert Lipsey and Helen Stone Tice, eds., *The Measurement of Saving, Investment and Wealth,* Chicago: University of Chicago Press, forthcoming 1988.

______, ______, O'Reilly, Terrance and Kumar, Praveen, "New Estimates of the Value of Federal Mineral Rights and Land," *American Economic Review*, December 1985, *75*, 923–36.

TABLE 1—VALUE OF FEDERAL OIL AND NATURAL GAS RIGHTS AND CHANGES IN VALUE 1954–86
(Billions of Current Dollars)

Year	Total	Oil	Gas	Change in Value
1954	58.3	46.5	11.8	
1955	58.3	46.3	12.0	0.0
1956	59.0	46.5	12.5	0.7
1957	64.5	51.4	13.1	5.5
1958	63.9	50.1	13.8	−0.6
1959	62.9	48.1	14.8	−1.0
1960	63.6	47.6	16.0	0.7
1961	64.9	47.7	17.2	1.3
1962	64.9	47.4	17.5	0.0
1963	64.9	47.2	17.7	0.0
1964	64.0	46.8	17.2	−0.9
1965	63.6	46.3	17.3	−0.4
1966	63.7	46.4	17.3	0.1
1967	64.0	46.6	17.4	0.3
1968	63.1	45.7	17.4	−0.9
1969	65.3	47.7	17.6	2.2
1970	65.8	48.2	17.6	0.5
1971	69.7	51.1	18.6	3.9
1972	67.4	49.2	18.2	−2.3
1973	74.0	54.0	20.0	6.6
1974	117.7	91.3	26.4	43.7
1975	138.8	100.7	38.1	21.1
1976	154.9	106.2	48.7	16.1
1977	174.6	109.2	65.4	19.8
1978	186.4	112.5	73.9	11.8
1979	247.2	153.2	94.0	60.8
1980	378.7	252.5	126.2	131.5
1981	521.2	362.8	158.4	142.5
1982	549.1	353.1	196.0	27.9
1983	532.6	324.9	207.7	−16.5
1984	520.3	316.6	203.7	−12.3
1985	491.5	288.9	202.6	−28.8
1986	334.8	172.2	162.6	−156.7

THE AMERICAN ECONOMIC REVIEW

VOLUME LXXVII

THE AMERICAN ECONOMIC ASSOCIATION

Executive Office: Nashville, Tennessee

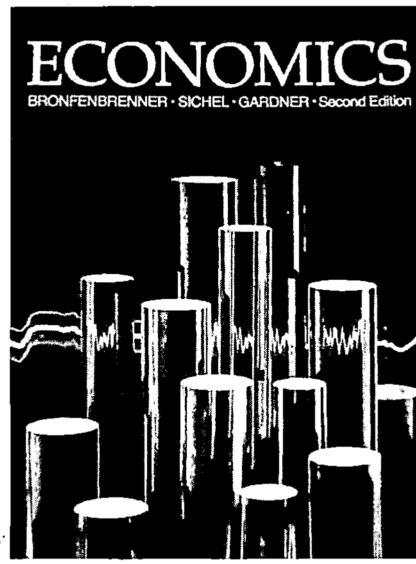
Editorial Office: 209 Nassau Street, Princeton, NJ 08542-4607

CONTENTS OF ARTICLES AND SHORTER PAPERS

CONTENTS OF THE PAPERS AND PROCEEDINGS

CONTRIBUTORS TO ARTICLES AND SHORTER PAPERS

CONTRIBUTORS TO PAPERS AND PROCEEDINGS


ECONOMICS
BRONFENBRENNER • SICHEL • GARDNER • Second Edition